MASTER'S THESIS

Hoe organisaties succesvol kunnen zijn in het managen van data science projecten

Deventer van, R.J. (Randy)

Award date:
2020

Hoe organisaties succesvol kunnen zijn in het managen van data science projecten

How organizations can be successful in managing data science projects

Opleiding:	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	R.J. van Deventer
Identiteitsnummer:	
Datum:	9 juli 2020
Afstudeerbegeleider	Prof.dr.ir R. Helms
Meelezer	Dr.ir. H.H. Martin
Versie nummer:	1.0
Status:	Definitief

Abstract

Dit onderzoek probeert meer inzicht te geven in welke type data science projecten er bestaan, welke methoden er gebruikt worden voor het uitvoeren van data science projecten en welke kritieke succesfactoren hierbij een rol spelen. Daarbij wordt op basis van de literatuurstudie een raamwerk voorgesteld waarbij er een relatie kan bestaan tussen het type project en de gebruikte methode. Daarnaast kan er een relatie bestaan tussen de methode en de kritieke succesfactoren die hierop van toepassing zijn. Het raamwerk heeft verder onderzoek nodig, maar kan een mechanisme zijn om data science projecten bewuster te sturen om een data science project succesvol op te leveren.

Sleutelbegrippen

Type en methode managen data science project, kritieke succesfactor

Samenvatting

De doelstelling van dit onderzoek is het creëren van meer inzicht in hoe data science projecten het beste kunnen worden uitgevoerd door kritieke succesfactoren te identificeren die invloed hebben op de uitvoer van het project, typen data science projecten in kaart te brengen en na te gaan hoe data science projecten gemanaged worden. Uiteindelijk moet dit inzicht leiden tot meer succes in het uitvoeren van data science projecten.

Op basis van de onderzoeksvragen is een literatuurstudie uitgevoerd. Het resultaat hiervan is dat er vier typen data science projecten te onderscheiden zijn en er drie methoden zijn die het meest gebruikt worden om deze projecten uit te voeren. Hierbij ligt in dit onderzoek de focus op CRISP-DM. Tenslotte zijn er 33 kritieke succesfactoren (KSF) vanuit de literatuur geselecteerd. Vervolgens is een raamwerk opgesteld dat een relatie legt tussen de type projecten en de methode om data science projecten te managen. Tevens is in het raamwerk een relatie te leggen tussen de methode om deze projecten te managen en de KSF'n die hiervoor van belang zijn.

In het empirisch deel van het onderzoek is middels case study bij twee organisaties het raamwerk getoetst. Hieruit komt naar voren dat er nog geen harde relatie te leggen is tussen het type data science project en de te gebruiken methode. Hetzelfde geldt voor de relatie tussen de methode en de KSF'n. Wel blijkt dat er zes KSF'n van belang worden geacht in een data science project dat gebruik maakt van CRISP-DM en de KSF die eruit springt is contact met de klant.
Organisaties delen hun data science projecten in naar een categorie in relatie tot hun business en niet in een bepaald type project. Tevens gebruiken ze CRISP-DM als handvat voor hun projecten en niet als strikte leidraad, waarbij CRISP-DM wordt aangevuld met best practices.

De beperking van dit onderzoek ligt in het beperkt aantal onderzochte organisaties.
Aanbevelingen voor verder onderzoek zijn dat typen data science projecten nader bekeken en geclassificeerd moeten worden. Daarnaast kan men op basis van dit onderzoek het raamwerk in een bredere omgeving toetsen. Ook zal de volwassenheid van organisaties met data science hierin meegenomen kunnen worden.

Inhoudsopgave

1. Introductie

1.1. Aanleiding

Dankzij technologische ontwikkelingen kunnen organisaties tegenwoordig grote hoeveelheden data verzamelen. Vervolgens willen deze organisaties de verzamelde data van waarde laten zijn. Redenen hiervoor zijn volgens Lavalle et al. (2011) het verbeteren van hun concurrentiepositie, het verbeteren van hun processen, kosten verminderen, sneller inspelen op de behoeftes van klanten, etc. Men wil als het ware in de toekomst kijken om hierop de strategische beslissingen te baseren, maar ook direct kunnen handelen om meteen bij te sturen. Een voorbeeld is de koffieketen Starbucks die succesvol data science gebruikte om een nieuwe smaak koffie te introduceren (Watson, 2014). Door op de dag van introductie de reacties op sociale media te monitoren kwam Starbucks erachter dat klanten de smaak van de koffie waardeerden, maar dat klanten de koffie ook te duur vonden. Hierop is direct gereageerd door de prijs te verlagen en de negatieve reacties verdwenen diezelfde dag op sociale media.
Om, net als bij Starbucks, de verzamelde data waardevol te laten zijn, worden door organisaties data science projecten gestart (Schüritz et al, 2017).

1.2. Probleemstelling

Data science omvat volgens Kellener (2018) een set beginselen, probleemdefinities, algoritmes en processen om vanuit grote datasets niet voor de hand liggende en bruikbare patronen te destilleren. Op basis hiervan kunnen organisaties inspelen op situaties en op verschillende niveaus beslissingen nemen. Ook het begrip big data past in deze context.
Helaas bereiken veel van deze projecten niet hun doel, waardoor organisaties niet kunnen profiteren van de verzamelde data (Gao, 2015). Uit onderzoek blijkt dat 55% (Kelly & Kaskade, 2013) tot zelfs 80% (Gartner, 2018) van de big data projecten falen en vele anderen hun doelen niet volledig halen. Concrete voorbeelden van gefaalde data science projecten bij organisaties komen niet naar buiten vanwege gezichtsverlies. Uiteindelijk kunnen mislukte data science projecten leiden tot hoge kosten, vertraagde planningen en worden de projecten aan de kant geschoven (Becker, 2017). Redenen hiervoor zijn dat de projecten geen bruikbare informatie opleveren, er geen duidelijke scope is of dat een organisatie simpelweg het project links laat liggen.

Probleemstelling:

Door organisaties opgezette data science projecten missen hun doel en leiden regelmatig tot niet succesvolle uitkomsten.

1.3. Opdrachtformulering

De doelstelling van het onderzoek is het creëren van meer inzicht in hoe data science projecten het beste kunnen worden uitgevoerd door kritieke succesfactoren te identificeren die invloed hebben op de uitvoer van het project, typen data science projecten in kaart te brengen en na te gaan hoe data science projecten gemanaged worden.

Onderzoeksvraag: Hoe kunnen data science projecten succesvol worden uitgevoerd door nader te kijken naar de typen projecten, methoden en kritieke succesfactoren?

Van de onderzoeksvraag worden de volgende deelvragen afgeleid:

1. Wat voor typen data science projecten kunnen worden onderscheiden?
2. Welke methoden zijn er beschikbaar voor het managen van data science projecten?
3. Welke kritieke succesfactoren zijn van invloed op data science projecten?

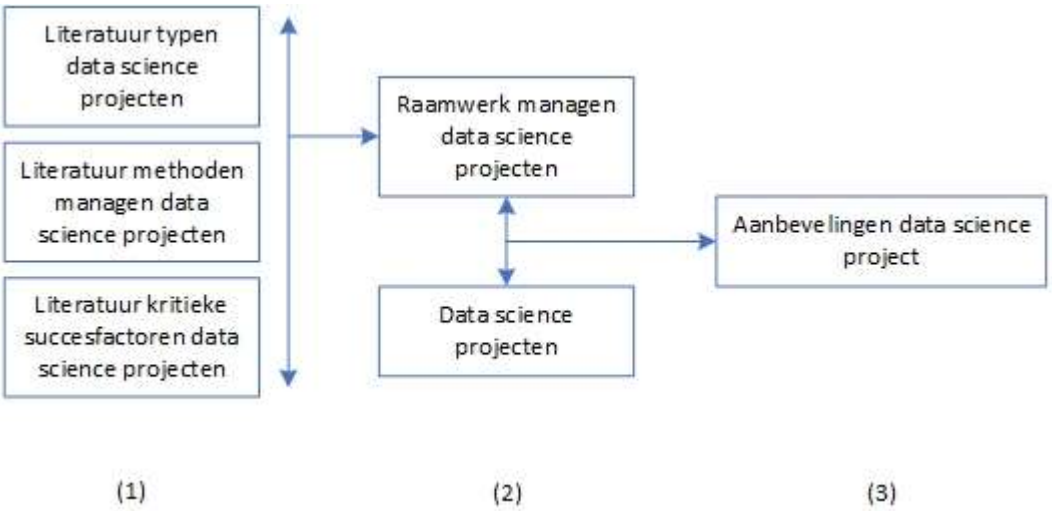
Om de onderzoeksvraag te beantwoorden wil ik via deelvraag 1 in de literatuur achterhalen welke typen data science projecten worden onderscheiden. Dit staat in relatie tot deelvraag 2 waarin wordt nagegaan welke methoden er onderkend zijn om data science projecten te managen. Door na te gaan welke typen data science projecten bestaan, kan worden bekeken of data science teams verschillende typen projecten anders behandelen door een specifieke methode te gebruiken om een bepaalde type data science project uit te voeren. In het empirisch deel kan onderzocht worden of er een relatie bestaat tussen een type data science project en de wijze van managen van het project en of dit van invloed is op het succesvol managen van data science projecten.
Vervolgens zal in deelvraag 3 bekeken worden op welke wijze een data science project succesvol kan zijn door in de literatuur na te gaan welke kritieke succesfactoren een data science project beïnvloeden. Waarna in het empirisch onderzoek wordt uitgezocht of bij organisaties deze kritieke succesfactoren een rol spelen bij data science projecten of dat andere factoren van invloed naar voren komen.

1.4. Relevantie

Het onderzoek moet duidelijk maken welke aspecten binnen een data science project van belang zijn om het data science project tot een goed einde te brengen, oftewel waardevol te zijn voor een organisatie. Er bestaat wetenschappelijke literatuur over hoe data science projecten verlopen, maar veel organisaties missen de ervaring en zijn nieuw in de wereld van data science projecten, naast de relatief snelle ontwikkelingen op dit gebied (Gao, 2015). Dat zou betekenen dat er behoefte is aan meer kennis voor organisaties om de kans van slagen van data science projecten te vergroten. Deze organisaties zijn dan wel in staat om winst te behalen in plaats van hoge kosten en onbruikbare data als resultaat. Daarnaast levert dit onderzoek een bijdrage aan de wetenschappelijke literatuur door kritieke succesfactoren te identificeren bij organisaties en deze factoren te toetsen. Tevens wordt nader gekeken naar typen data science projecten. Dit vult bestaand onderzoek aan met betrekking tot data science projecten en factoren van invloed hierop en geeft nieuwe inzichten.

1.5. Onderzoeksaanpak

Dit onderzoek zal kwalitatief van aard zijn. In de wetenschappelijke literatuur zal worden gezocht naar eerdere onderzoeken waarin is gebleken welke factoren van invloed zijn op data science projecten. In eerste aanzet zal er worden gezocht naar relevante literatuur van de laatste tien jaar, omdat in deze periode data science een snelle ontwikkeling doormaakt.
Tijdens het empirisch onderzoek zal door middel van semigestructureerde interviews bij drie organisaties worden nagegaan welke kritieke succesfactoren een rol hebben gespeeld tijdens het data science project.



Figuur 1 - Onderzoeksmodel

Het onderzoeksmodel (Verschuren, 2015) beschrijft (1) een bestudering van de wetenschappelijke theorie over typen data science projecten, methoden om data science projecten te managen en kritieke succesfactoren van invloed op deze projecten, dat leidt tot een raamwerk van kritieke succesfactoren in relatie tot typen en management methodes van data science projecten, (2) waarmee door interviews bij drie organisaties die data science projecten uitvoeren het raamwerk getoetst/aangevuld wordt. (3) De uitkomst leidt tot aanbevelingen die data science projecten een grotere kans van slagen geven.

2. Theoretisch kader

2.1. Onderzoeksaanpak

Het doel van dit theoretisch kader is om een raamwerk samen te stellen waardoor inzicht ontstaat in hoe data science projecten succesvol gemanaged kunnen worden. De methode waarop de literatuurstudie gebaseerd is om het theoretisch kader te bouwen, is de "systematic literature review of information systems research" afkomstig van Okoli & Schabram (2010). In deze methode worden acht stappen beschreven om op systematische wijze een literatuurstudie uit te voeren. Het gaat om de volgende stappen:

1. Doel van de literatuurstudie, om aan te geven wat bereikt moet worden.
2. Protocol en training, om bij een onderzoek met meerdere reviewers de wijze van werken af te stemmen en hen hierin te trainen.
3. Zoeken van de literatuur door het selecteren van de relevante artikelen.
4. Praktische filtering van artikelen door gebruik van een aantal criteria.
5. Beoordeling van de kwaliteit (kwalitatief) door na te gaan hoe de argumentatie in het artikel is.
6. Data extractie, oftewel de data die nodig is voor de antwoorden op de onderzoeksvragen selecteren.
7. Het analyseren en samenvoegen van de data.
8. Het schrijven van de literatuurstudie met de antwoorden op de onderzoeksvragen (in par. 2.3)

In dit onderzoek wordt stap 2 (protocol en training) achterwege gelaten, omdat er geen gebruik gemaakt wordt van meer dan één reviewer.

2.2. Uitvoering

Er wordt gestart met stap 3, het zoeken van de literatuur. De wetenschappelijke artikelen zullen gezocht worden met behulp van de zoekmachines van de bibliotheek van de Open Universiteit[1] en Google Scholar[2]. Beide zoekmachines hebben connecties met een groot aantal academische databases, maar ze verschillen in het gebruik van zoektermen door andere operators. Hierbij wordt gezocht naar artikelen in de Engelse taal.
De term "data science" wordt voornamelijk de laatste jaren gebruikt, maar eerder werd de term "big data" al veel gebruikt.

Hieronder wordt per vraag aangegeven welke zoektermen gebruikt worden. Het kan voorkomen dat het gevonden artikel niet relevant is voor bijvoorbeeld vraag 1, maar wel voor vraag 2, ondanks dat de zoekterm voor vraag 1 is gespecificeerd. In het overzicht in bijlage 1 wordt aangegeven voor welke vraag een gevonden artikel dient. In bijlage 1 wordt tevens aangegeven welke query gebruikt is. Om de zoekresultaten te beperken wordt er alleen naar trefwoorden in de titel gezocht.

Voor het beantwoorden van vraag 1 wordt gezocht naar de volgende termen of een combinatie hiervan:
"data science projects", "big data projects", "type".

[1] https://bibliotheek.ou.nl
[2] https://scholar.google.com

Bij vraag 2 worden de volgende termen of een combinatie hiervan gebruikt:
“methods data science projects”, “managing data science projects”, “manage data science projects”, “methods big data projects”, “managing big data projects”, “manage big data projects”, “data science projects”, “big data projects”, “methods”, “managing”, “manage”, “methodologies”, “methodology”.

En tenslotte bij vraag 3 de volgende termen:
“critical success factors”, “data science projects”, “big data projects”, “data science”, “big data”.

Door de gebruikte query’s leverden de zoekopdrachten via de online Open Universiteit Bibliotheek relevante, maar beperkte zoekresultaten op. Via Google Scholar werden door gebruik van iets andere query’s meer resultaten gevonden, maar de meest relevante artikelen werden bovenaan de lijst met resultaten gevonden. Uiteindelijk bleken de gebruikte query’s resultaten te produceren met dezelfde terugkerende artikelen waarna het zoeken naar nieuwe literatuur is gestopt. De gevonden artikelen zijn geregistreerd en opgeslagen via de applicatie Endnote X9[4]. In de tabellen 4 en 5 in bijlage 1 staan de gebruikte query’s en zoekresultaten van respectievelijk de Open Universiteit Bibliotheek en Google Scholar.
Als blijkt dat het artikel relevant lijkt te zijn, dan zal in een volgende stap de inhoud van het artikel gescand worden.

De volgende stap (4) is het praktisch filteren van de artikelen door gebruik te maken van een aantal criteria. Hierbij wordt gelet op de gebruikte taal in het artikel (Engels) en er wordt gezocht naar relevante artikelen in de laatste 15 jaar, dus vanaf 2004 tot en met 2019. Daarnaast wordt gekeken naar de wijze van publicatie (in een gerenommeerd wetenschappelijk tijdschrift, conference paper) en de auteur (als blijkt dat deze in meerdere publicaties over dit onderwerp terugkomt, aantal keer geciteerd volgens Google Scholar). De door de zoekmachines gepresenteerde artikelen zullen eerst op de (context van de) titel en abstract gescand worden. Hierbij wordt gelet of de titel en abstract de zoektermen bevatten, maar bij de abstract wordt ook naar het doel van het artikel gekeken en of dat aansluit bij de deelvragen in dit onderzoek. De typen artikelen die gezocht worden zullen peer reviewed zijn. Tevens kan het voorkomen dat in artikelen verwijzingen naar andere artikelen staan met betrekking tot dit onderwerp en die eventueel relevant blijken te zijn voor het beantwoorden van de onderzoeksvragen als ze aan de criteria voldoen, de zogenaamde sneeuwbalmethode. In bijlage 2 tabel 6 staat de uiteindelijke selectie van de artikelen die gebruikt worden voor het theoretisch kader.

In bijlage 3 tabel 7 bevinden zich de opmerkingen over de artikelen die behoren bij stap 5, het beoordelen van de kwaliteit. Hierbij is gekeken naar hoe er in de artikelen beweringen en conclusies zijn beargumenteerd. Bijvoorbeeld wanneer een theorie/model meerdere keren getoetst is, de gebruikte methode is beschreven en de situatie waarin is onderzocht, dan zal het artikel van goede kwaliteit zijn.

Het resultaat van de stappen 6 tot en met 8 staat in paragraaf 2.3 en zal de antwoorden geven op de onderzoeksvragen en leiden tot een raamwerk voor verder onderzoek.

[4] https://endnote.com

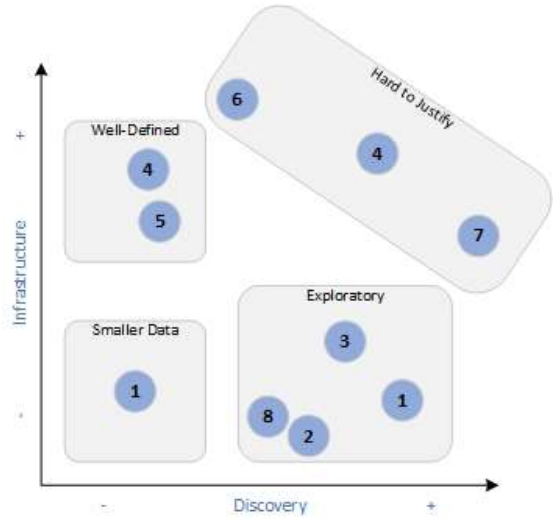
2.3. Resultaten en conclusies

2.3.1. Typen data science projecten

Twee typen data science projecten die worden onderscheiden zijn “routine projecten” en “verkennende projecten” (Saltz & Shamshurin, 2015). Routine projecten worden op regelmatige basis uitgevoerd, hebben een standaard proces en een goed beschreven methodologie. Dit geobserveerde project bleek maar één echte fase van een data science project te bevatten (preprocessing), waarbij de auteurs concluderen dat dit type project niet als een data science project beschouwd kan worden. Data science projecten die een erkende en beschreven methode volgen (zie volgende paragraaf) bestaan uit meerdere fases. In de verkennende projecten komen wel de fases naar voren die in een data science project gebruikt worden (voorverwerking, data analyse en implementatie), maar gebruiken geen standaard methodologie en hebben geen vaste tijdsduur. Ook niet per fase van een project. Aangezien dit onderzoek bij maar één organisatie gedaan is op basis van observaties en het onderzoek zich verder richtte op het verkennende project, is de wetenschappelijke theorie van de twee onderkende typen zwak. In het artikel wordt tevens opgemerkt dat de mate van volwassenheid van de organisatie met data science projecten laag en lastig te documenteren is.

Een ander onderscheid van type projecten dat in de literatuur wordt gemaakt is: “moeilijk te rechtvaardigen”, “verkennend”, “duidelijk gedefinieerd” en “weinig data” (Saltz et al, 2017). In dit artikel is, ten opzichte van het eerste artikel, grondiger onderzoek gedaan op basis van 14 karakteristieken van een data science project. Het raamwerk dat hieruit volgde concentreert zich op twee karakteristieken: ontdekking en IT infrastructuur. Waarbij de mate van ontdekking bepaald wordt door twee factoren en mate van infrastructuur bepaald wordt door vijf factoren. In een 2D diagram worden ontdekking en infrastructuur geplot, zodat er in feite vier kwadranten ontstaan die de vier typen projecten onderscheiden:

- Moeilijk te rechtvaardigen: project heeft geen helder doel, maar vereist vooraf een grote investering. Moeilijk om ondersteuning vanuit de organisatie te krijgen.
- Verkennend: project heeft geen helder doel, dus makkelijker om zaken te proberen. Lage kosten door minder vereiste infrastructuur. Dit project werd met deze benaming ook beschreven in het eerste artikel en komen overeen op het volgende punt: geen helder doel.
- Duidelijk gedefinieerd: project heeft een duidelijk doel, maar vergt een grote investering. Vooraf is namelijk te rechtvaardigen dat de investering nut heeft.
- Weinig data: project met een duidelijk doel, maar vergt een kleine investering in infrastructuur.



Figuur 2 – Vier typen data science projecten (Saltz et al, 2017)

Als basis voor dit onderzoek wordt uitgegaan van de vier typen projecten die Saltz et al (2017) beschreven hebben. Reden hiervoor is het gestructureerde onderzoek en ook hun aanbeveling (in de inleiding) dat op basis van deze vier typen data science projecten in verder onderzoek kan worden bepaald welke project methodologie hierop het beste aansluit.

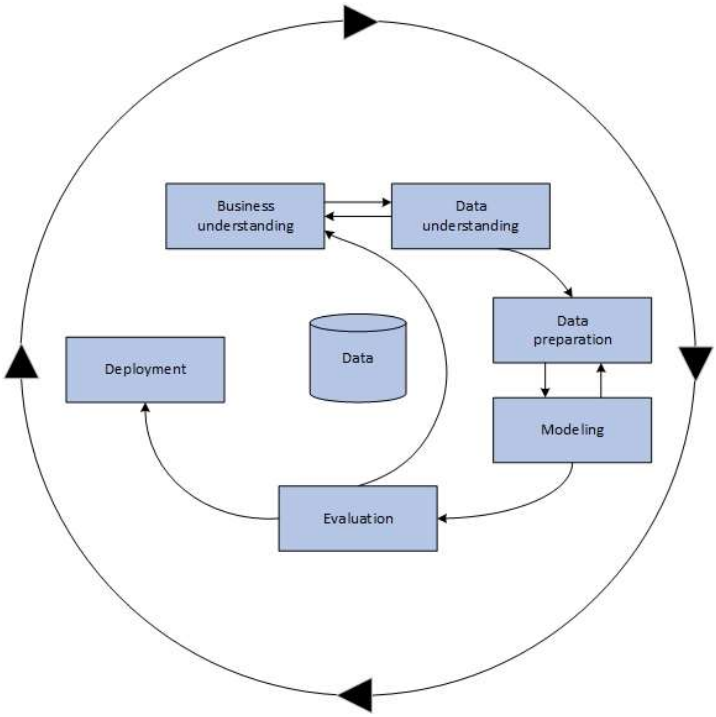
2.3.2. Methoden voor managen data science projecten

In deze paragraaf wordt antwoord gegeven op de vraag welke methoden er beschikbaar zijn voor het managen van data science projecten.

In de literatuur zijn meerdere artikelen verschenen die methoden beschrijven om data science projecten uit te voeren. Vaak wordt ook gesproken over big data, knowledge discovery in databases (KDD) en data mining (Mariscal et al, 2010). In de literatuur worden meerdere methoden beschreven en aanpassingen/verbeteringen op deze methoden. Op hoofdlijnen lijken de methoden op elkaar, maar op detailniveau verschillen ze van elkaar (Saltz et al, 2017). Vanwege de beschikbare tijd voor dit onderzoek, beperk ik me tot het beschrijven van de drie meest voorkomende en bewezen methoden voor data science projecten.

CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) wordt onder andere door Mariscal et al (2010), Saltz et al (2017), Schmidt et al (2018) als de meest gebruikte methode voor data science. Het procesmodel bestaat uit 6 fases: business understanding (doelen en eisen vanuit de business), data understanding (verzamelen van en inzicht in data), data preparation (dataset voorbereiden), modeling (algoritmen selecteren en toepassen), evaluation (evalueren model of het de doelen haalt) en deployment (uitvoeren van het model en presenteren van de data). Deze fases zijn onderverdeeld in taken en processen bij die taken. CRISP-DM is industrie-, toolset- en applicatieneutraal. Sinds 2007 bestaat er ook een versie CRISP-DM 2.0.



Figuur 3 – CRISP-DM (Chapman et al, 2000)

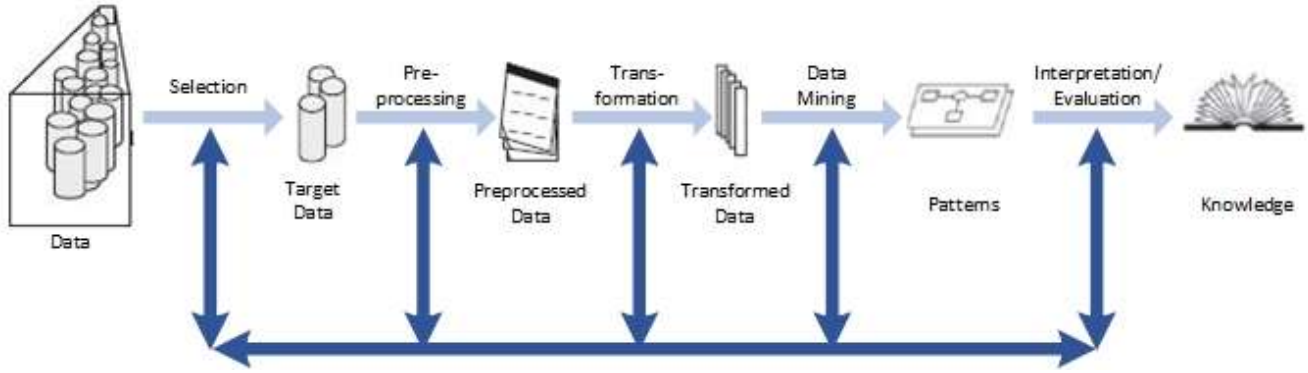
SEMMA

SEMMA is ontwikkeld door het SAS Institute en staat voor de stappen die het doorloopt tijdens het data mining proces, namelijk Sample, Explore, Modify, Model en Assess. Deze methode is volgens Saltz et al (2017) en Dutta (2015) de tweede meest gebruikte methode en is afgeleid van het uitgebreide KDD model uit 1996. Hierbij mist SEMMA twee stappen t.o.v. het KDD model hieronder die essentieel zijn om een data mining proces succesvol uit te voeren. Daarnaast maakt SEMMA deel uit van een toolset, is het geen open proces en kan het niet in elke omgeving gebruikt worden (Mariscal et al, 2010).

Figuur 4 – SEMMA proces (SAS Institute, 2005)

KDD

Het KDD proces is door Fayyad in 1996 gepubliceerd met als doel patronen in data te identificeren en te gebruiken. Het bestaat uit negen stappen: het toepassingsdomein leren kennen, een data set creëren, data opschonen en voorbereiden, data reductie en projectie, functie kiezen voor data mining, algoritme bepalen, data mining, interpretatie patronen en visualisatie en gebruiken van de opgedane kennis (Mariscal et al, 2010). Dit is het derde meest gebruikte proces voor data science projecten (Schmidt et al (2018), Saltz et al (2017), Rogalewicz (2016)



Figuur 5 – KDD (Fayyad et al, 1996)

Methode	Fases					
CRISP-DM	Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
SEMMA		Sample	Explore	Model	Assess	
		Explore	Modify	Assess		
KDD	Learning the application domain	Creating a target data set	Data cleaning and pre-processing	Choosing the function of DM	Interpretation	Using discovered knowledge
			Data reduction and projection	Choosing the DM algorithm		
				Data mining		

Tabel 1 – Overzicht methoden (Mariscal et al, 2010)

Conclusie

Voor dit onderzoek wordt de focus gelegd op CRISP-DM als de te gebruiken methode voor het managen van een data science project. Redenen hiervoor zijn dat CRISP-DM de meest gebruikte methode is en dit de kans vergroot op het vinden van een organisatie voor de case study. Daarnaast is de methode onafhankelijk van de bedrijfstak waarin deze gebruikt wordt en zijn er geen specifieke applicaties of toolsets voor nodig. Dit zal tevens de scope van het onderzoek smaller maken.

2.3.3. Kritieke succesfactoren en data science projecten

Welke KSF'n van invloed kunnen zijn op data science projecten wordt in deze paragraaf beschreven. Voor de term kritieke succesfactor wordt de volgende definitie gegeven: "*the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization. They are the few key areas where 'things must go right' for the business to flourish.*" (Rockart, 1979). Een andere definitie die gegeven wordt is dat *een kritieke succesfactor een methode, management tool, of ontwerptechniek is die een effectieve ontwikkeling en uitvoering van een proces of project mogelijk maakt* (Gartner, 2019) . Kortom een factor die op verschillende gebieden in een project en organisatie kan voorkomen, maar tevens noodzakelijk is om een project succesvol te laten zijn.

Er zijn een groot aantal KSF'n in relatie tot het uitvoeren van data science projecten geïdentificeerd in de verschillende artikelen, waarvan er een deel met elkaar overeen komen. De wijze waarop de KSF'n zijn genoemd, gerangschikt per categorie of zijn afgeleid, verschillen van elkaar. In deze paragraaf zal ik een enkele lijst van KSF'n selecteren.

Gao et al (2015) borduurt verder op het werk van Koronios et al (2014) met in totaal 27 KSF'n. De KSF'n zijn in tabel 8 in bijlage 4 te vinden. Er is door hen gezocht naar KSF'n op drie gebieden, namelijk mensen, proces en technologie. Deze drie gebieden zorgen voor succes of falen in een IT-project (Sicular, 2012). Vervolgens zijn de KSF'n ingedeeld in de fase waarin ze van belang zijn voor een big data project. De zes fases in het procesmodel vertonen gelijkenissen met de zes fases uit tabel 8. Eybers et al (2017) hebben een zelfde indeling in drie categorieën opgezet op basis van verschillende onderzoeken. Zij noemen de categorieën mens, organisatie (incl proces) en technologie.

Daarnaast is er een lijst met 33 KSF'n geïdentificeerd en deze KSF'n zijn ingedeeld volgens zes karakteristieken die een organisatie, die volwassen is op het gebied van big data projecten, kenmerkt (Saltz & Shamshurin, 2016). Deze lijst is een uitbreiding op de lijst van Gao et al (2015). De KSF'n staan in tabel 2 hieronder.

Voor het raamwerk zal worden uitgegaan van de lijst met KSF'n opgesteld door Saltz & Shamshurin (2016) in tabel 2, omdat dit de meest complete lijst is. Zoals Saltz en Shamshurin aangeven is op basis van deze lijst verder onderzoek nodig om de KSF'n te priotiseren, verfijnen en valideren. De categorieën voor de KSF'n, die in de eerdergenoemde artikelen zijn gebruikt, worden losgelaten. Reden hiervoor is dat onderzocht gaat worden welke KSF'n relevant zijn voor een bepaalde methode om een data science project succesvol te managen.

Nr	KSF en *categorie*
	Data
1	Data & data quality management / ownership
2	Data integration & security
3	Unstructured/structured data
4	Representativeness of data
5	Document collection/access to sources
	Governance
6	Management priority / sponsorship / support
7	Big data strategy alignment (with organization's vision)
8	Project management process defined
9	Well defined organizational structure
10	Performance management
11	Data protection and privacy by design
12	Culture of being data-driven
	Process
13	Close collaboration between IT and business
14	Communication about the data and initiatives
15	Flexibility and agility, with freedom for experimentation
16	Focus on change management
17	Project difficulty explored and communicated
18	Clarity of project deliverables (clear or ambiguous)

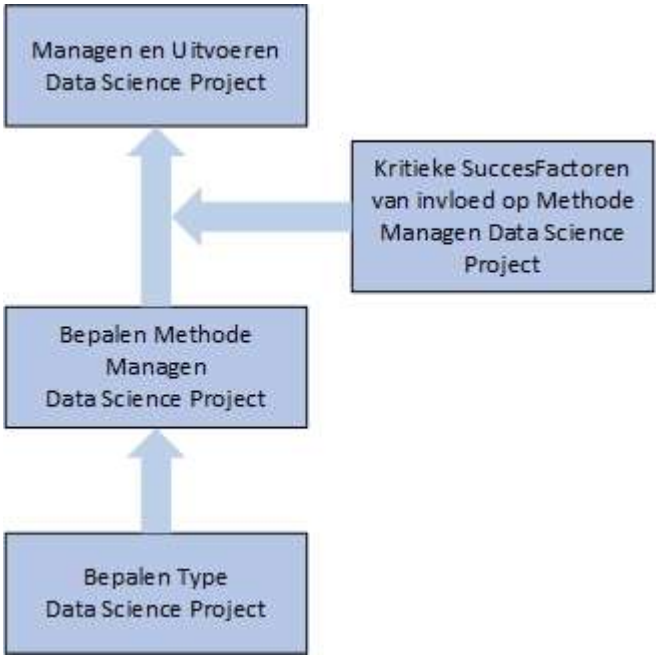
Nr	KSF en *categorie*
18	Clarity of project deliverables (clear or ambiguous)
	Objectives
19	Focus on small projects and known questions
20	Specified business case
21	Feasibility study
22	Skill gap analysis
23	Well defined scope - that understood by the team
24	Measurable project outcome
	Team
25	Development of skills / training
26	People skills & ability to self-organize when needed
27	Data science, technology, business & management skills
28	Multidisciplinary team (i.e., across different departments)
29	Stakeholder coordination / shared understanding
	Tools
30	Investment in IT infrastructure, technology & tools
31	Investment in data sources & data storage
32	Reporting and visualization technology
33	Discovery technology

Tabel 2 – KSF Saltz & Shamshurin (2016)

2.3.4. Conclusie

Er zijn vier typen data science projecten te onderscheiden afhankelijk van welke IT infrastructuur benodigd is en in welke mate er data onderzocht/ontdekt moet worden. Voor het bepalen van de methode om data science projecten uit te voeren, wordt uitgegaan van het meest gebruikte processen, namelijk CRISP-DM. Op basis van de karakteristieken van deze processen kan worden bekeken welke methode het beste aansluit op welk type data science project.
De lijst met kritieke succesfactoren kan vervolgens worden gerelateerd aan de methode die toegepast wordt voor het managen van het data science project.
Dit geheel wordt voorgesteld in het raamwerk in figuur 6 en zal worden getoetst en uitgewerkt in het empirisch deel van het onderzoek middels de volgende twee proposities:

1. Het type data science project bepaalt de methode die gebruikt wordt om een data science project te managen.
2. Een methode om een data science project te managen vereist specifieke KSF'n.



Figuur 6 – Raamwerk managen data science project

3. Methodologie

3.1. Keuze van de onderzoeksmethode

Het doel van het empirische onderzoek is het valideren of bijstellen van het raamwerk. De relevantie hiervan is dat er in de wetenschappelijke literatuur geen raamwerk bestaat dat aangeeft welke methode om data science projecten te managen het beste aansluit bij een bepaald type data science project. Waarbij ook de relevantie van KSF'n is meegenomen bij het managen van een data science project.

Om het onderzoeksdoel te bereiken dient te worden nagegaan op welke wijze organisaties de data science projecten uitvoeren, of de type data science projecten hierin terugkomen en of er een relatie hiertussen bestaat. Daarnaast moet worden bekeken of de geïdentificeerde KSF'n uit het raamwerk bestaan en van invloed zijn bij de data science projecten van de organisaties.

Hiervoor moet onderzoek worden gedaan bij organisaties die data science projecten uitvoeren binnen hun organisatie (binnen een project) of in opdracht van een organisatie.

Het onderzoek is kwalitatief en verklarend van aard, omdat er gezocht wordt naar de relaties tussen variabelen (type data science project en methode managen data science project) en waarom/hoe die relatie bestaat. Er zijn verschillende methoden om informatie voor dit onderzoek te vinden.

Saunders (2016) beschrijft een aantal methoden om deze informatie te verkrijgen. De methode die voor dit onderzoek gebruikt zal worden is de "case study" en in de volgende paragraaf wordt de motivatie hiervoor uiteengezet.

3.2. Case study

Een case study is een diepgaand onderzoek in een onderwerp of fenomeen binnen een levensechte omgeving, waarbij de case een organisatie kan zijn (Yin, 2014). Hierbij is het van belang dat in het onderzoek de context waarin het fenomeen wordt bestudeerd helder is. Er worden door Yin vier typen case studies beschreven in een matrix: enkele case t.o.v. meerdere cases en holistisch t.o.v. ingebed.

Voordelen:
- Een case study is uitermate geschikt om een theorie te ontwikkelen door een fenomeen in een levensechte omgeving te bestuderen.
- Een case study kan gedetailleerde of diepgaande data opleveren die een andere methode niet kan opleveren.

Nadelen:
- De verzamelde data bij een case study is misschien niet representatief voor een bredere populatie als er een beperkt aantal cases wordt onderzocht, waardoor de relevantie afneemt (externe validiteit).
- De interne validiteit kan afnemen doordat een enkele onderzoeker data subjectief interpreteert.

De case study is in dit geval de beste onderzoeksmethode voor dit kwalitatieve en verklarende onderzoek, specifiek gezien voor dit onderzoek een holistische meervoudige case study. Het te onderzoeken fenomeen (case) is het managen van een data science project. Om na te gaan hoe dit project wordt uitgevoerd en een relatie te kunnen leggen tussen de variabelen (type, methode en KSF bij een data science project) is diepgaand onderzoek nodig. De uitkomsten hiervan kunnen bijdragen aan een theorie die het managen van data science projecten succesvol maakt. Een

nadelige factor is de beperkte tijd, waardoor maar een beperkt aantal cases kan worden onderzocht. Mocht er relevante secundaire data (documentatie) bij de organisatie aanwezig zijn, dan wordt dit meegenomen in het onderzoek.

3.3. Technisch ontwerp van onderzoek

De wijze waarop de case study zal worden uitgevoerd, zal middels semigestructureerde interviews zijn. Reden hiervoor is het kwalitatieve en verklarende karakter van het onderzoek. Hierbij is het mogelijk bepaalde thema's (type, methode en KSF) met een aantal basisvragen te onderzoeken en vervolgens hierop dieper in te gaan middels doorvragen. Dit kan mogelijk per organisatie verschillen, maar het doel is dit zo consistent mogelijk uit te voeren.

Voor de interviews zijn er gericht drie organisaties als primaire bron uitgezocht die zich bezig houden met data science projecten, waarbij die organisaties beschikken over een speciaal daarvoor ingerichte data science afdeling. Deze organisaties kwamen in beeld door contacten binnen de organisatie waar de onderzoeker werkzaam is. De te interviewen functionarissen dienen deel uit te maken van een data science team, zodat ze over voldoende kennis beschikken om de vragen te beantwoorden. Bij voorkeur worden meerdere functionarissen binnen het team geïnterviewd om een bredere dataverzameling te verkrijgen.
Het interview zal starten met een aantal opmerkingen om het doel van het onderzoek helder uit te leggen en de context te schetsen. Dit kan aanknopingspunten voor het gesprek bieden. In bijlage 5 staat het schema voor de werkwijze voor het interview.
De data wordt tijdens het interview verzameld door aantekeningen te maken en door van het interview een audio opname te maken met goedkeuring van de geïnterviewde.

In de interviews is het nodig per thema of deelvraag de volgende gegevens te verzamelen:

1. Welke typen data science projecten kunnen worden onderscheiden in de organisatie?
    a. Nagaan of de organisatie verschillende typen data science projecten onderkent of onderscheid en waarom.
    b. Zo niet,
        i. Nagaan of de organisatie de vier typen data science projecten herkent die beschreven zijn in de literatuur.
        ii. Zijn de data science projecten van de organisatie toe te passen in de matrix uit de literatuur?
2. Welke methoden zijn er beschikbaar voor het managen van data science projecten?
    a. Nagaan welke methode(n) de organisatie gebruikt voor het managen van data science projecten.
    b. Nagaan welke reden er is voor het gebruik van deze methoden.
    c. Nagaan of de organisatie bekend is met de twee methoden die beschreven zijn in de literatuur.
    d. Bekijken of er een relatie tussen het type project en de methode bestaat.
3. Welke kritieke succesfactoren zijn van invloed op data science projecten?
    a. Nagaan of de organisatie (bewust) gebruik maakt van KSF'n (welke en waarom wel/niet).
    b. Zo niet, zijn er KSF'n te herkennen in de projecten?
    c. Bekijken of er een relatie bestaat tussen de methode en de KSF'n.

In bijlage 5 is de vragenlijst voor de interviews opgenomen.

3.4. Gegevensanalyse

De verzamelde data is kwalitatief van aard en verzameld middels semigestructureerde interviews. Van de audio opname van het interview zal een verslag gemaakt worden. Het verslag zal zo snel mogelijk na het interview geschreven moeten worden om associaties met andere interviews te voorkomen. Dit verslag zal door de geïnterviewde nagelezen worden en eventueel becommentarieerd worden, zodat diegene er zeker van is dat de uitgeschreven uitspraken correct zijn. Een nadeel hiervan is dat de geïnterviewde het verslag kan “verbeteren”, waardoor de uitkomst van het gesprek anders wordt.

De te gebruiken analysemethode is template analyse (Saunders et al, 2014). Deze methode is systematisch, flexibel en relatief eenvoudig toe te passen op kwalitatieve data ongeacht de invalshoek (deductief en inductief). Dat is een voordeel als een onderzoeker weinig ervaring heeft met wetenschappelijk onderzoek en het geeft de mogelijkheid om de data op een gestructureerde manier te analyseren middels een template. Op deze wijze kan er meer tijd worden gestoken in een grondig onderzoek in plaats van tijd te verspillen door regels van een complexe methode toe te passen.

De template analyse volgt de volgende stappen:

3. Bekend worden met de data door van de interviews verslagen te maken.
4. Code template opstellen met behulp van de onderwerpen uit het interviewprotocol.
5. Coderen van de verslagen door de stappen, beschreven door Ose (2016), te volgen. Op deze manier kunnen kwalitatieve onderzoeksresultaten op systematische wijze gecodeerd worden. Tijdens het coderen de template bijstellen door het toevoegen, samenvoegen of wijzigen van codes.
6. Zoeken van de thema’s en relaties herkennen.
7. Het toetsen van onderstaande proposities:
    a. Het type data science project bepaalt de methode die gebruikt wordt om een data science project te managen.
    b. Een methode om een data science project te managen vereist specifieke KSF’n.

3.5. Reflectie t.a.v. validiteit, betrouwbaarheid en ethische aspecten

Om het onderzoek verantwoord op te zetten, zal rekening moeten worden gehouden met de volgende criteria. Als eerste de *bias*, waarbij de antwoorden van de geïnterviewde worden genoteerd in een verslag en teruggekoppeld aan de geïnterviewde, zodat de interpretatie van de onderzoeker een zo laag mogelijke invloed heeft op de data. Het gevaar bestaat wel dat de geïnterviewde geen volledig antwoord durft te geven vanwege vertrouwelijkheid. Ten tweede de *betrouwbaarheid*, die in dit onderzoek is beschreven volgens welke methode het onderzoek wordt uitgevoerd en waarom, hoe de data wordt verzameld en wordt geanalyseerd. Hiermee wordt het proces dat wordt gevolgd duidelijk voor andere onderzoekers, zodat zij dit kunnen gebruiken voor eigen onderzoek. Het nadeel is dat de case study een momentopname is en dat de situatie van het onderzochte thema kan veranderen. Daarnaast moet rekening worden gehouden met de *generaliseerbaarheid*. Door het beperkt aantal cases van maximaal 3 stuks bestaat de kans dat de generaliseerbaarheid laag is. In dit onderzoek wordt geprobeerd meerdere functionarissen binnen het data science team te ondervragen, zodat er een bredere dataverzameling is. Daarnaast is de overdraagbaarheid van het onderzoek hierboven aangehaald bij betrouwbaarheid. Als laatste speelt de *validiteit* een rol. Criteria voor validiteit (geloofwaardigheid, overdraagbaarheid) zijn hierboven

aangehaald. Het kan voorkomen dat lopende het onderzoek er zaken bijgesteld worden, waarbij de kans bestaat dat de zaken die bijgesteld worden niet voldoende worden bijgehouden.

Om de onderzoeksethiek te waarborgen, zal er rekening worden gehouden met de volgende uitgangspunten. Vanwege onderzoek bij een overheids- en commerciële organisatie wordt de verkregen data vertrouwelijk behandeld. Tevens wordt er integer met de identiteit van de betrokkenen bij de organisatie en met de organisatie zelf omgegaan door deze anoniem te beschrijven in het onderzoek. Daarnaast wordt het onderzoek bij de organisatie alleen gedaan op basis van goedkeuring door de organisatie en toestemming van de organisatie om de verkregen data te gebruiken voor het onderzoek. Tenslotte wordt bij de te onderzoeken organisatie open en eerlijk aangegeven wat het doel van het onderzoek is en hoe er met de verkregen data wordt omgegaan. Dit wordt middels een consentformulier gedaan dat door de onderzoeker en de te interviewen medewerker kan worden getekend.

4. Resultaten

4.1. Uitvoering interviews

Voor het onderzoek was het de bedoeling bij elk van de drie organisaties drie interviews te houden. Uiteindelijk bleek dat het door omstandigheden rondom het COVID-19 virus moeizaam was om medewerking te vinden van organisaties en zijn er interviews niet doorgegaan. De interviews hebben vanwege richtlijnen rondom het virus op afstand plaatsgevonden via Skype in plaats van de geplande locatie bij de organisatie. Van de interviews zijn met toestemming opnames gemaakt voor verwerking ervan tot een verslag.

Uiteindelijk zijn er vier interviews doorgegaan, waarbij twee respondenten ook ervaringen hebben gedeeld van hun werkzaamheden bij data science projecten van hun eerdere werkgever. Na het interview is het opgemaakte verslag (nagenoeg een transcriptie) naar de respondent gestuurd ter review. Tevens heeft de respondent het KSF-formulier met opmerkingen geretourneerd. Hierna is gestart met de template analyse van de onderzoeksresultaten.

4.2. Bevindingen per organisatie

4.2.1. Organisatie 1

De eerste organisatie is een overheidsorganisatie bestaande uit ongeveer een paar duizend medewerkers. Sinds 2018 beschikt deze organisatie over een data science afdeling die producten levert aan klanten/gebruikers binnen de organisatie. Deze producten moeten zorgen voor winst, maar niet direct op een financiële manier. De winst richt zich op dit moment met name op het efficiënter inrichten van processen waardoor uiteindelijk wel kosten kunnen worden bespaard. Daarnaast is het doel van de data science afdeling is het halen van meerwaarde uit data door van die rauwe data informatie te maken die te gebruiken is bij de besluitvorming van de organisatie. Voorbeeld van een project zijn prognoses maken voor het komende jaar hoeveel medewerkers moeten worden aangenomen in een bepaalde vakgroep en op welk niveau op basis van de historische data. Ook wordt een planning voor de inzet van systemen gemaakt op basis van beschikbare data (variabelen). Het maken van die planning kost veel manuren en middels dit project wordt geprobeerd dat efficiënter te maken. De geïnterviewde medewerker is een data analist die werkzaam is binnen de expertisegroep modelleren.

De data science afdeling bestaat uit ongeveer 30 medewerkers en is nog steeds in ontwikkeling. Hierbij wordt men ondersteund door een externe organisatie. De afdeling is op twee manieren ingedeeld: volgens een product/business owner en volgens expertisegroepen. Het product owner deel geeft aan in welke categorie een data science project valt, namelijk Maintenance, Communicatie & Beïnvloeding, Personeelslogistieke keten en Operatiën. De categorie van het project is afhankelijk van de klantvraag of het probleem dat voor een klant moet worden opgelost. De product owner kan vervolgens beschikken over vier expertise teams om het project te doorlopen. Dat zijn Pipeline, Modelleren, Inzicht en Front-end. Deze expertise teams komen gedurende een project niet in een vaste volgorde binnen een project aan de orde. Afhankelijk van waar behoefte aan is op een bepaald moment binnen een project, wordt een expertise team ingezet. Het kan voorkomen dat het project van het ene expertise team naar het andere expertise team gaat.

Een data science project start met een klantvraag die helder moet zijn voor het team. Het team krijgt vervolgens toegang tot de benodigde data en bouwt daar een tool voor, zodat deze de vraag van de klant kan beantwoorden. De gebruikte methode voor het uitvoeren van projecten is CRISP-DM voor

het volgen van het proces en scrum voor het ontwikkelen van het product. Elk project doorloopt dat proces, maar hier hangt geen vaste tijdsduur aan vast. De sprints duren 4 weken en op basis van de review en retro wordt de voortgang van het project bewaakt en bijgestuurd. Er wordt bijvoorbeeld 2 dagen per week aan een project besteed, zodat er meerdere projecten tegelijkertijd uitgevoerd kunnen worden. Elk project wordt een user story genoemd en de user story bevindt zich telkens in een bepaald gedeelte van dat CRISP-DM proces. Dat proces stuurt het team van waar het zich in het proces bevindt en welke vervolgstappen nodig zijn. Dat wordt vertaalt in (kleine) issues, die per persoon worden opgelost. Na het oplossen van een x aantal issues kom het team weer in een volgende stap. . Het is een iteratief proces, maar als de wens van de klant is ingevuld, wordt het project afgesloten.

De reden voor het gebruik van CRISP-DM is dat het een bewezen methode is en data science voor de organisatie nieuw is. Daarbij wordt CRISP-DM niet alleen voor het hele project gebruikt, maar ook voor de deelproducten in een project.

Eindverantwoordelijkheid voor het project ligt bij de scrummaster (hoofd van de afdeling), maar deze delegeert de dagelijkse verantwoordelijkheid naar de product owner voor communicatie met de klant, prioriteiten stellen en het opleveren van de (deel)producten. Het team onder leiding van de product owner voor een project is daarbij zelfsturend binnen de randvoorwaarden van de scrummaster. Dat team bestaat uit de eerdergenoemde expertisegroepen, waarbij een bepaalde expertise afhankelijk van de opdracht de nadruk heeft in een project.

4.2.2. Organisatie 2

De tweede organisatie is tevens een overheidsorganisatie, maar specifiek gericht op het leveren van IT diensten binnen een departement. De IT diensten worden aan verschillende organisaties binnen de overheid geleverd. In 2017 is het data science team opgestart en het bestaat inmiddels uit 20 medewerkers en is groeiende. Oorspronkelijk was de afdeling gestart als een kenniscentrum voor data science binnen de organisatie. De afdeling bestaat uit een mix van projectmanagers, data scientists en data engineers. De geïnterviewde medewerkers zijn de oprichter/programmamanager van het team, die zich ook bezig houdt met de besturing en data governance, en twee data scientists. Voorbeelden van projecten zijn prognoses voor onderhoud van systemen binnen de organisatie opstellen en voorspellen van geomagnetische stormen voor beschikbaarheid van GPS. Daarnaast wordt er binnen de organisatie infrastructuur beschikbaar gesteld om data science producten uit te voeren. Ook hier is het doel van data science om processen in de organisatie efficiënter te laten verlopen en om beter en sneller in staat zijn beslissingen te nemen.

In de meeste gevallen start een data science project met een opdracht vanuit een klant. Uitzondering is als er wordt onderkend dat er op een bepaald gebied kennis mist voor de data science afdeling. Op basis van een interview met de klant over doel, urgentie, beschikbaarheid van data, etc. wordt een plan van aanpak opgesteld. Dit plan van aanpak bevat de opdracht, de gewenste output, tijdlijn met milestones, wanneer het project eventueel stopt, hoe men data aangeleverd krijgt, etc. Op basis van urgentie en beschikbare capaciteit wordt vervolgens bepaald welk project wordt uitgevoerd. In de toekomst gaat hier de regiegroep van de CIO een rol in spelen.

Als het project in uitvoering komt, wordt gebruik gemaakt van CRISP-DM waarbij ook scrum en best practices van PRINCE-2 gebruikt worden. Daarbij wordt de klant betrokken in de vorm van verwachtingsmanagement. Met de klant wordt overlegd over de implementatievorm en een proof of concept (POC), waar na verschillende iteraties het POC in productie kan worden genomen en de lifecycle start. Tijdens de lifecycle zorgt het team ervoor dat het product blijft functioneren door

bijvoorbeeld algoritmes bij te stellen als dat nodig is. Het gehele proces is beschreven in een "levend" document en wordt bijgewerkt op basis van ervaringen die opgedaan worden met huidige projecten of door (bestuurlijke) randvoorwaarden die door de organisatie opgelegd worden. Het data science team dat aan een opdracht werkt is wisselend, afhankelijk van de opdracht en de grootte ervan. Het bestaat uit een projectleider, een materiedeskundige, data scientist en eventuele overige teamleden. Er is een gedeelde verantwoordelijkheid voor een project binnen het team tijdens de uitvoering, maar de projectmanager is wel degene die het proces bewaakt. De verantwoordelijkheid verschuift naar de "keten" binnen de organisatie zodra het product geoperationaliseerd/geïmplementeerd wordt. De keten kan worden gezien als een divisie binnen de organisatie.

In tabel 3 hieronder is een overzicht van de onderzochte organisaties en de eigenschappen ervan bij data science projecten te vinden.

Org	Indeling projecten	Methode	Initiatie	Verantwoordelijkheid
1	In 4 categorieën (per product owner)	CRISP-DM met scrum	Opdrachtgever	Scrummaster delegeert aan product owner Expertise teams met eigen verantwoordelijkheid
2	Per thema/functie van product	CRISP-DM met scrum en PRINCE-2	Opdrachtgever en regiegroep	Gedeelde verantwoordelijkheid van team Projectmanager coördineert

Org	Ervaring	Teamsamenstelling project	Sturen op KSF	KSF van belang
1	2 jaar	Product owner met vertegenwoordiging 4 expertise teams	Nee	Klantcontact
2	3 jaar	Mix van project manager, data scientists, data engineers, materiedeskundige	Nee	Klantcontact

Tabel 3 – Overzicht eigenschappen organisaties met data science projecten

4.3. Resultaten op de onderzoeksvragen

4.3.1. Wat voor typen data science projecten kunnen worden onderscheiden?

In organisatie 1 worden projecten ingedeeld volgens categorieën van de product owners, namelijk Maintenance, Communicatie & Beïnvloeding, Personeelslogistieke keten en Operatiën. De categorie van een project zegt niks over hoe een project uitgevoerd moet worden, maar alleen voor welk doel het resultaat of product van een project gebruikt gaat worden in de organisatie. De data scientist geeft aan dat "De projecten zou ik exact indelen in de eerder genoemde categorieën van product owners". Reden hiervoor is dat de verbinding naar de business helder is.
De in het onderzoek voorgestelde typen van data science projecten worden niet gebruikt, maar wel herkend. Zo is het mogelijk de door organisatie 1 uitgevoerde projecten in te delen in de matrix, maar wordt de matrix beperkt gevonden. Een voorbeeld is het personeelsprognose project. Volgens de geïnterviewde "is er voldoende rekenkracht en er is voldoende data, maar voor de data is nog extra werk nodig om deze te gebruiken voor een prognose. Het project past in het exploratory kwadrant". Een ander voorbeeld is volgens hem het project jaarplanning systemen dat "in het kwadrant well-defined zou passen, want de data is redelijk compleet en er zijn goede bewezen methodes/algoritmes, waardoor je snel een goede uitkomst kan krijgen". Door de data scientist wordt geen alternatief gegeven om projecten volgens andere criteria in te delen.

Organisatie 2 deelt de projecten in per functie of thema waar het op te leveren product in valt. Zo geeft de programmamanager aan dat “De projecten worden gecategoriseerd volgens de thema’s/functie van de opdrachten, bv. maintenance, targeting, etc. Het doel is om business alignment te krijgen tussen mensen die in organisatietermen denken en de data scientist die in algoritmes/datasets denken.”
De voorgestelde matrix uit de literatuur wordt niet direct als bruikbaar gezien, al worden de typen wel herkend. Een data scientist gaf aan dat “een slimme assistent voor call center medewerkers. Dat project zat rechtsonder, dus veel discovery en weinig infrastructure. Maar dat had ook te maken met de fase van het project, omdat dat meer een proof of concept was. Als je de POC zou willen uitbreiden, dan ga je wel veel infrastructuur nodig hebben om het uit te rollen in de organisatie. Een project zou kunnen verschuiven binnen de matrix”. Een andere data scientist zegt dat “De meeste projecten bij het datalab vallen in de exploratory categorie, omdat er veel uitgezocht en gemaakt moet worden”. Dit laatste heeft tevens te maken met de oorsprong van het data science team als kenniscentrum voor data science en het innovatieve karakter.
De matrix van Saltz et al (2017) zou aangepast kunnen worden. Volgens de geïnterviewde zou “het model meer dimensionaal moeten zijn. Infrastructuur, maar ook de reden/doel. Dan zijn er nog steeds projecten die “hard to justify” zijn zoals in het model. Dat zijn zaken die in de toekomst liggen, waarvan mensen nog niet de relatie naar de business toe begrijpen” en “Het onderzoekende (exploratory) kent ook meer dimensies richting de tools”. Daarnaast geeft hij aan dat “Er moeten meer dimensies aan het model worden toegevoegd. Hierbij denk ik aan besturing (roadmapping, grootte van projecten), onderzoekend, infrastructuur, business, operationaliseren.”
Er zijn meerdere invalshoeken bij het beantwoorden van deze vraag waar te nemen. Zo wordt geopperd om projecten in te delen op basis van de soort data die je gebruikt: “Je zou projecten dus kunnen indelen in: gestructureerde data, tekst data, beeld data, geo data”. Ook kan men kijken naar “Enerzijds technisch (engineering, analytisch, ontsluiting, visualisatie). Anderzijds ook wel in een bepaalde volwassenheidsfase [mate van innovatie]”.
In verder onderzoek zou een nieuw raamwerk voor type data science projecten kunnen worden opgezet met meerdere lagen of dimensies. Hierbij zet ik als kanttekening neer dat een dergelijk raamwerk niet te complex mag zijn om het acceptabel en bruikbaar te maken voor toepassing door data science teams.

Beide organisaties geven aan dat de data science projecten worden ingedeeld op het thema of de categorie waarin het product valt en niet op een type project. Reden hiervoor is dat de organisaties dit beter vinden aansluiten op de business. Deze indeling staat los van een type project zoals beschreven door Saltz et al (2017) waarin het gaat om bepaalde eigenschappen van het project (infrastructuur en mate van ontdekking).

4.3.2. Welke methoden zijn er beschikbaar voor het managen van data science projecten?

Organisatie 1 gebruikt CRISP-DM voor het proces bij de uitvoer van een project en scrum in de verschillende iteraties voor het ontwikkelen van het product. In de uitvoering zijn de zes fases van CRISP-DM waar te nemen. Zo beschrijft de data scientist het verloop van een project “dat begint bij een klantvraag die we proberen te begrijpen” (*business understanding*). Bij de *data understanding* “krijgen we de data om de klantvraag te kunnen beantwoorden”, waarbij men met behulp van de klant meer inzicht in de data probeert te krijgen. *Data preparation* werd door de geïnterviewde niet letterlijk uitgelegd. Hij geeft wel aan dat bijvoorbeeld in het personeelsprognose project, dat er data moet worden voorbereid middels “er is voldoende data, maar voor de data is nog extra werk nodig

om deze te gebruiken voor een prognose". Vervolgens start *modeling* en "het data science team bouwt daar iets voor, zodat de klant daar zijn vraag mee kan beantwoorden". Tijdens de *evaluation* bekijkt men de volgende aspecten in het lopende project: "op basis van de review en retro bewaken we de voortgang van de projecten, wat is er goed of niet goed gegaan, wat is er af en niet af en wat willen we meer of minder in het project." Tenslotte volgt de *deployment* fase, waarbij het product aan de klant geleverd wordt en dat gebeurt "als het naar wens van de klant is ingevuld, dan is het project klaar."

Ook organisatie 2 gebruikt CRISP-DM voor het uitvoeren van een project. Het proces wordt beschreven in een "levend" document aangezien de organisatie op het gebied van data science aan het groeien is. Voor CRISP-DM geeft men daarin aan dat "Het proces van data science is een cyclus. Deze cyclus bestaat uit zes fases" en "De volgorde van deze fases is niet vast; er zit een wisselwerking tussen". Uit het gesprek komt naar voren dat CRISP-DM gebruikt wordt, maar "dat zegt niks over de besturing. Er wordt tevens gebruik gemaakt van de goede dingen uit PRINCE-2 op een flexibele manier". En er wordt nog een stap verder gegaan door te stellen dat "de methode is eigenlijk niet van belang. Er zijn verschillende methodes en als je deze methodes naast elkaar legt, lijken ze op elkaar. Het is normaal projectmanagement dat is getailleerd op data science". Dit is in lijn met de literatuur aangezien Mariscal et al (2010) met hun overzicht van procesmodellen laten zien dat onder andere CRISP-DM afstamt van het KDD proces uit 1993. Daarnaast wordt scrum gebruikt in projecten, waarbij de data scientist aangeeft "dat proberen we ook via scrum uit te voeren, maar als je met zijn tweeën bent, kan het flexibeler. Het is een hele intuïtieve methode van korte sprintjes, even plannen wat je nu doet en wat de volgende stap gaat zijn, omdat dat fijn werkt". Hier laat men zien dat men niet strak vasthoudt aan een methode, maar per project of situatie inspeelt op het gebruiken van een methode.
In de fase van *business understanding* wordt "gezocht naar de motivatie van het project en wordt samen met de klant naar oplossingsrichtingen gezocht. Op basis van de beperkingen, middelen en organisatiecultuur, wordt besloten welke aanpak het beste past (hoe)". De *data understanding* fase heeft als doel meer inzicht te krijgen in de data door "een evaluatie van de verzamelde gegevens op kwaliteit en kwantiteit" en "een fundamentele analyse om samenhang tussen de verstrekte gegevens en het doel (van de klant) te bevestigen". Tijdens de *data preparation* fase wordt de te gebruiken data bepaald aan de hand van criteria als "relevantie voor de doelen, kwaliteit van de gegevens en technische beperkingen". *Modeling* wordt gestart door "het selecteren van een geschikte techniek, bijvoorbeeld tijdreeksanalyse of cross-sectie analyse" en "maken we een mechanisme om de kwaliteit en de validiteit van ons model te testen". Bij de *evaluation* fase wordt nagegaan of het model "voldoet aan de inhoudelijke doelstellingen en controleren we of dit model effectief en efficiënt is". Als laatste fase volgt deployment waarin het product wordt geïmplementeerd. Hiervoor "is er vaak een pilot nodig om de resultaten te laten landen in uw organisatie. Als de pilot is geslaagd worden de resultaten geoperationaliseerd".

Hieruit blijkt dat CRISP-DM door de onderzochte organisaties als handvat gevolgd wordt tijdens het data science project en niet strikt als leidend geheel voor het project. Tevens maakt men zoals aangegeven in paragraaf 4.1.1 en 4.1.2 gebruik van "hulpmiddelen" zoals PRINCE-2 en best practices.

4.3.3. Welke kritieke succesfactoren zijn van invloed op data science projecten?

Organisatie 1 stuurt niet bewust met behulp van KSF'n in projecten. Er zijn wel KSF'n waar te nemen bij de uitvoering van projecten. De belangrijkste KSF is contact met de klant en deze regelmatig

betrekken tijdens het proces, zodat telkens getoetst wordt of hetgeen je bouwt het product is dat de klant verwacht. Daarnaast is in relatie tot deze KSF een heldere opdracht van de klant van belang. Ook worden betrouwbare data en de beschikbaarheid van voldoende technische infrastructuur nodig geacht als KSF. Dit laatste aspect ging mis bij een project voor de planning van systemen. Hierbij gaf de geïnterviewde aan dat "het oplossen van het lineair probleem is computerintensief en het kwam wel eens voor dat de het proces bleef hangen, waardoor er een reset nodig was van de hardware. Dat kan negatieve effecten hebben op andere projecten. Of er gaat iets mis in het begin van de rekenintensieve taak, wat je achteraf pas ziet en dat kost tijd".

Organisatie 2 stuurt niet altijd op KSF'n. Op basis van de interviews komen verschillende reacties. Er wordt gesteld dat er gestuurd wordt "op basis van data, domeinkennis en opdrachtgeverschap", maar dat dit niet consequent gebeurd : "Er is geen vast patroon om KSF te gebruiken. De ene keer wordt wel een datakwaliteitsrapport voor de data opgemaakt, maar als de hoeveelheid data klein is bijvoorbeeld niet". Ook is niet altijd de kennis aanwezig om op KSF'n te sturen, omdat "Dit is lastig te beantwoorden, omdat ik nooit les gehad heb in wat deze kritieke succesfactoren zijn". Een reden hiervoor is tevens dat deze data science afdeling nog vrij jong is. Er wordt niet altijd (bewust) gestuurd op KSF'n, maar men gebruikt wel de ervaring en best practices en dus onbewust een KSF. Dat blijkt voor de KSF klantcontact uit "Wat wel actief wordt gedaan in het proces is dat je de opdrachtgever hierin meeneemt en hem continu op de hoogte houdt, constant vraagt om feedback te geven en bij te sturen waar nodig".

Op basis van de lijst van Saltz & Shamshurin (2016) hebben de respondenten de tien belangrijkste KSF'n voor uitvoering van een project geselecteerd. Hier zijn zes KSF'n uit naar voren gekomen die het meest van belang zijn op basis van het aantal keer dat ze zijn aangevinkt door de respondenten. Het gaat om:

1. *Data & data quality management / ownership*: het is van belang dat men tijdig over de data kan beschikken. Tevens dient de kwaliteit/kwantiteit van de data goed te zijn om geen vreemde uitkomsten uit het product te krijgen. Bij organisatie 2 liep men bij het PNOD project tegen dit probleem aan, "omdat er geen data werd aangeleverd en via de opdrachtgever kwam men ook niet bij de data. Meestal gaat een project fout door de data: het niet hebben van data, maar ook slechte kwaliteit van data". Door regelmatig contact met de eigenaar van de data kan er gewerkt worden aan data van betere kwaliteit, maar ook dat er altijd juiste data beschikbaar is.
2. *Representativeness of data*: de data moet de juiste weergave zijn van wat geanalyseerd moet worden, dus geen synthetische data. Zonder de juiste data is de praktische toepassing van het model beperkt. Een data scientist onderbouwt dit met "in sommige sectoren is momenteel nog geen goede data voor handen om mee te experimenteren, dan moet je naar oplossingen zoals synthetische data gaan kijken, maar in mijn ervaring is dat nog nooit goed uitgepakt.
3. *Management priority / sponsorship / support*: belangrijk voor bestaansrecht binnen een organisatie, maar ook als escalatiemogelijkheid bij problemen in de organisatie. Het hogere management zorgt voor budget om de projecten te kunnen uitvoeren en vervolgens het product te implementeren. Zonder deze ondersteuning slaagt geen enkel project door gebrek aan draagkracht. Dat blijkt uit een antwoord van een data scientist uit organisatie 2 bij een matching project toen men inzag dat een project waarde had: "Dat heeft heel erg geholpen om de business case nog sterker te krijgen om meer te investeren om het project verder te brengen." Er zal dus ook over projecten en de waarde hiervan naar het hoger management gecommuniceerd moeten worden.
4. *Culture of being data-driven*: een organisatie moet transformeren naar een data gedreven organisatie en hier het belang van inzien. Door deze basis ontstaat de situatie dat projecten succesvoller zijn, doordat meer data beschikbaar en dat de data van betere kwaliteit is, omdat

men zich hier van bewust is. Een voorbeeld waarmee organisatie 2 te maken kreeg is dat er nieuwe voertuigen gekocht werden en “bij inkopers moet op het netvlies staan dat bij inkoop van nieuwe voertuigen ook toegang tot de sensordata meegekocht moet worden”. En dit voorbeeld wordt nog breder in de organisatie weggezet door aan te geven dat “bij degenen die het voertuig besturen moet bekend zijn dat die data belangrijk is, en dat sensoren dus niet zomaar uitgezet, verwijderd, etc. moeten worden”.

5. *Close collaboration between IT and business*: zeer regelmatig contact tussen de klant en het data science team zorgt voor een heldere opdracht en een product dat de klant accepteert, draagvlak bij de eenieder, het vergroot de kans dat het product waarde toevoegt, verwachtingsmanagement. In organisatie 1 kwam dit naar voren bij het project voor de planning van systemen, waarbij de data scientist aangaf dat klantcontact nodig is om “om er achter te komen wat het probleem exact is. Daarnaast is het van belang om de klant constant te betrekken bij het project om te bekijken of hetgeen wat je bouwt is wat de klant verwacht”.
6. *Clarity of project deliverables* (clear or ambiguous): door heldere en behapbare doelen te stellen (en hierbij de klant te betrekken) voor een project weet iedereen in het team wat er bereikt moet worden om een succesvol eindproduct te leveren. Een voorbeeld hiervan is het project om de optimale route van een schip bepalen op basis van verschillende invalshoeken in organisatie 2. Dit project was succesvol, omdat de data scientist aangaf: “ging volgens mijn inschatting het contact met de business heel goed” en “want als je niet dezelfde doelen voor ogen hebt, gaat je product nooit op een goede manier in ontvangst genomen worden”. De andere data scientist laat weten “dat was het continu afstemmen met de eindgebruiker. Dat is uiteindelijk wel waarvoor je het doet. Het echte succes van een project wordt gedefinieerd of je het product gebruikt of niet, of het werkt en toegevoegde waarde heeft”.

De onderzochte organisaties sturen niet op KSF’n, maar ze komen al dan niet bewust terug in de uitvoering van data science projecten. Dat kan te maken hebben met de volwassenheid van de organisatie in de uitvoering van deze projecten, maar die factor is niet meegenomen in dit onderzoek. Uiteindelijk zijn er zes KSF’n die men het meest belangrijk vindt, al sprong tijdens de interviews de regelmatige communicatie met de klant als KSF er uit.

4.3.4. Toetsing proposities

Er zijn twee proposities gedaan, die tijdens het empirisch onderzoek werden getoetst om te bekijken of er een relatie tussen deze factoren te vinden was:

1. Het type data science project bepaalt de methode die gebruikt wordt om een data science project te managen.
2. Een methode om een data science project te managen vereist specifieke KSF’n.

De eerste propositie is in dit onderzoek niet bewezen. Redenen hiervoor zijn dat men in de onderzochte organisaties projecten niet indeelt in bepaalde typen, maar bijvoorbeeld wel in een bepaalde categorie vanwege de indeling van de data science afdeling of op basis van het product dat geleverd moet worden (bv. Maintenance of Personeelslogistiek in organisatie 1). Tevens is er nog niet veel bekend over typen data science projecten aangezien er in de literatuur alleen in het artikel van Saltz (2017) onderzoek naar gedaan is. Vanuit de interviews is voorgesteld het model uit te breiden met extra dimensies of lagen, maar men gaf ook aan dat sommige projecten in het model van Saltz (2017) zijn toe te passen. Ook gebruikten de data science teams in dit onderzoek een vaste methode, namelijk CRISP-DM, en wisselden ze niet van methode. Hierdoor zijn er geen verschillende methoden om data science projecten te managen in het empirisch deel onderzocht. Dat heeft te maken met de beperkte scope van het onderzoek.

Hetzelfde geldt nagenoeg ook voor de tweede propositie in dit onderzoek die niet hard bewezen kan worden. De data science teams stuurden niet bewust met KSF'n om projecten succesvol te laten zijn. De teams leggen deels en onbewust de nadruk op bepaalde KSF'n om een project succesvol te doorlopen. Zo geeft men in organisatie 2 aan drie KSF'n te gebruiken, maar ook dat er niet bewust gebruik wordt gemaakt van KSF'n. Het blijkt wel dat KSF'n waar te nemen zijn, maar niet dat deze KSF'n specifiek voor deze methode om data science projecten te managen het beste te gebruiken zijn. Er zijn zes KSF'n naar voren gekomen die door deze organisaties van belang worden geacht voor het uitvoeren van de projecten waarbij ze CRISP-DM gebruiken. Maar het kan ook zijn dat deze zes KSF'n voldoen voor andere methoden om deze projecten uit te voeren. Zoals eerder aangegeven is in dit onderzoek alleen CRISP-DM meegenomen. In verder onderzoek zou voor meerdere methoden bekeken moeten worden welke KSF'n per methode van invloed zijn op een succesvol project.

5. Discussie, conclusies en aanbevelingen

5.1. Discussie – reflectie

In dit onderzoek was het aantal organisaties en respondenten beperkt door omstandigheden (in de wereld, namelijk Covid-19). Het vinden van organisaties die wilden meewerken aan het onderzoek was lastig. Dit heeft als gevolg dat er geen breed beeld van verschillende organisaties die data science projecten uitvoeren is ontstaan. De uitkomsten van te gebruiken methoden en de motivatie hiervoor om deze projecten uit te voeren is beperkt. Dit vervolgens tussen meerdere organisaties vergelijken is ten dele gelukt. Ondanks deze beperking is het mogelijk om conclusies uit het onderzoek te trekken die leiden naar aanbevelingen voor verder onderzoek. Zo is wel duidelijk geworden welke KSF'n relevanter worden gevonden om een data science project succesvol tot een einde te brengen. Deze KSF'n zijn niet nieuw ten opzichte van de uitgebreide lijst KSF'n uit de literatuur.

De impact van de beschikbare tijd voor het onderzoek is groot. Het onderzoek werd in deeltijd naast de reguliere baan uitgevoerd, waarbij de laatste prioriteit kreeg. Dat heeft als gevolg gehad dat ik minder de diepte in ben gegaan in het onderzoek.

De scope van het onderzoek was in het begin lastig te bepalen. Door gebruik te maken van een startset van wetenschappelijke literatuur en rekening te houden met het oogmerk van het onderzoek (hoe worden data science projecten gemanaged), is hier richting aan te geven. Toch wil je een bijdrage leveren aan wetenschappelijk onderzoek en ben je zoekende naar iets nieuws waar meer duidelijkheid over gegeven zou kunnen worden om data science projecten succesvol te maken. Dat is gevonden in het nagaan welke typen data science projecten bestaan, de methoden om deze projecten te managen en de KSF'n die hierbij een rol kunnen spelen. Vervolgens is er gezocht naar een relatie tussen deze onderwerpen. Dat resulteerde in een raamwerk dat nog niet in de literatuur bestond, maar dat raamwerk heeft verder onderzoek nodig om de voorgestelde relaties nader te bekijken. Reden hiervoor ligt in het beperkte aantal organisaties waar onderzoek is gedaan. De typen data science projecten uit de literatuur waren deels te herkennen bij de organisaties, maar werden als zodanig niet gebruikt om een project uit te voeren.
Het onderzoek was waarschijnlijk te breed. Achteraf gezien had er beter geconcentreerd kunnen worden op een relatie tussen typen projecten en KSF'n of de relatie tussen methoden en projecten. De scope had dus wat beperkt moeten worden gezien de beschikbare tijd.

5.2. Conclusies

Het blijkt dat de typen projecten uit het model van Saltz et al (2017) niet eenduidig door de respondenten te herkennen zijn. De typen projecten die in de literatuur naar voren komen zijn beperkt en gebaseerd op een matrix van "hoeveelheid infrastructuur" en "de mate van ontdekking" in een project waardoor 4 typen projecten zijn te onderscheiden. De projecten indelen volgens een type om vervolgens hierop te sturen, komt niet voor bij deze organisaties. Dit kan te maken hebben met de mate van volwassenheid met data science projecten bij deze organisaties, omdat de onderzochte organisaties maximaal 3 jaar bestaan. Toch worden projecten ingedeeld in een categorie, maar dat staat los van hoe men een data science project aanloopt. De categorie van het project heeft te maken met de klantvraag of het product dat geleverd moet worden.

In de wetenschappelijke literatuur komt naar voren dat er een aantal methoden het meest gebruikt worden voor het uitvoeren van data science projecten. Dat zijn CRISP-DM, SEMMA en KDD. In grote lijnen beschikken deze methoden over dezelfde 6 fases en verschillen deze methoden op het gebied

van toepasbaarheid. De gebruikte methode bij de onderzochte organisaties is CRISP-DM. Reden hiervoor is dat dit een bewezen methode is die structuur biedt bij het uitvoeren van data science projecten. De methode is niet leidend tijdens de uitvoering van een project, maar biedt voornamelijk houvast. Men wijkt soms af van CRISP-DM, omdat dat vanwege gezond verstand, ervaring of best practices (van PRINCE-2) in het project beter werkt. Men denkt hiermee betere resultaten te halen. Dat betekent dat de organisaties wel de fases van CRISP-DM onderkennen, maar niet persé aan deze volgorde vasthouden.

Er zijn een groot aantal KSF'n waar rekening mee gehouden kan worden of waarmee men tijdens een data science project kan sturen. In de literatuur blijft het bij een lange lijst aan KSF'n die geïdentificeerd zijn in eerder onderzoek, maar er worden geen KSF'n aangegeven die voor het meeste succes kunnen zorgen. De KSF die in de praktijk het meest naar voren kwam is klantcontact, oftewel zeer regelmatige communicatie tussen de klant en de product owner van het data science team.

Er is geen duidelijke relatie tussen een type project en de te gebruiken methode om het project te managen in het onderzoek naar voren gekomen. Dat betekent overigens niet dat die relatie niet bestaat, omdat het aantal organisaties in het onderzoek zeer beperkt was. Daarbij wordt ook opgemerkt dat de typen projecten uit het raamwerk waarschijnlijk niet volledig zijn.

Tevens is er geen duidelijke relatie tussen de te gebruiken methode en de te gebruiken KSF'n voor het managen van een data science project. Ook hier is niet met zekerheid te zeggen dat deze relatie niet bestaat vanwege de beperkte diepte van het onderzoek. Er werd door deze organisaties namelijk niet bewust gebruik gemaakt van KSF'n om te sturen op een succesvolle uitkomst, maar wel bleken een aantal KSF'n gedurende het project van belang te zijn.

5.3. Aanbevelingen voor de praktijk

Het blijkt dat men in de praktijk een bekende methode voor het managen van een data science project vooral gebruikt als handvat, maar niet als verplicht te volgen leidraad. Dit kan afhankelijk zijn van de grootte van het project, de ervaring met data science projecten, de teamleden en hun ervaring binnen een project, etc. Hierbij kan men gebruik maken van best practices of een methode aanvullen met een projectmanagementmethode zoals PRINCE-2.

De KSF die als belangrijkste in het onderzoek naar voren kwam is regelmatig klantcontact oftewel "close collaboration between IT and business". Een data science project is succesvol als de klant tevreden is met het opgeleverde product. Dat betekent dat het product aan de wensen van de klant moet voldoen, bruikbaar is en het gewenste doel bereikt. Dat kan alleen als de klant nauw betrokken wordt bij het proces en een rol speelt (feedback geeft) bij de iteraties van de producten.

5.4. Aanbevelingen voor verder onderzoek

Saltz et al (2017) geven aan dat er meer onderzoek moet worden gedaan naar of hun raamwerk voldoet voor gebruik door data science teams. Daar is in dit onderzoek een start mee gemaakt. Er moet worden bekeken of dit raamwerk voorzien kan worden van meer dimensies of invalshoeken in plaats van de huidige twee (infrastructuur en verkenning) dimensies. Dat kan leiden tot meerdere typen projecten, maar het zal tevens getoetst moeten worden of ze toepasbaar zijn in meerdere werkdomeinen. In de huidige literatuur is over type data science projecten weinig te vinden.

Vervolgens kan worden onderzocht of er specifiek een relatie bestaat tussen het type project en de methode om een data science project te managen. Het is raadzaam het onderzoek niet te breed te

maken, maar dit per onderzoek te focussen op een specifieke methode zoals CRISP-DM. Daarnaast dient er per methode onderzocht te worden welke KSF'n positief of negatief van invloed zijn op het verloop van het project.

Een belangrijke factor die niet meegenomen is in dit onderzoek, maar wel door de respondenten aangehaald werd, is de mate van volwassenheid van een organisatie met data science projecten. Bij volgend onderzoek moet deze factor worden meegenomen om uitkomsten in beter perspectief te plaatsen. De volwassenheid met data science projecten kan wat zeggen over het toepassen van de methode om data science projecten te managen, de mate waarin men al dan niet bewust KSF'n gebruikt en de mate waarin men succesvol is in data science projecten.

Referenties

Becker, D. (2017). *Predicting outcomes for big data projects: Big Data Project Dynamics (BDPD): Research in progress.* Paper presented at the 2017 IEEE International Conference on Big Data, Boston, MA, USA.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. Retrieved from

Dutta, D., Bose, I. (2015). Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics, 165*, 293-306. doi:10.1016/j.ijpe.2014.12.032

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). *Knowledge discovery and data mining: Towards a unifying framework.* Paper presented at the KDD-96 Proceedings.

Gao, J., Koronios, A., Selle, S. (2015). *Towards a process view on critical success factors in big data analytics projects.* Paper presented at the Twenty-first Americas Conference on Information Systems, Puerto Rico.

Gartner. (2018). Gartner says nearly half of CIOs are planning to deploy artificial intelligence.

Gartner. (November 2019). Csf (Critical Success Factor). Retrieved from https://www.gartner.com/en/information-technology/glossary/csf-critical-success-factor

Institute, S. (December 2019). Introduction to SEMMA. Retrieved from https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

Kellener, J. D., Tierney, B. (2018). *Data science*. Cambridge, MA: The MIT Press.

Kelly, J., Kaskade, J. (2013). CIOs & BIG DATA What your IT team wants you to know.

Koronios, A., Gao, J, Selle, S. (2014). *Big data project success - a meta analysis.* Paper presented at the PACIS 2014 Proceedings.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 20 - 31.

Mariscal, G., Marbán, Ó., Fernández, C. . (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25:2*, 137-166.

Okoli, C., Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Sprouts: Working Papers on Information Systems, 10*(26).

Ose, S. O. (2016). Using Excel and Word to structure qualitative data. *Journal of Applied Social Science*, 1 - 16.

Rockart, J. F. (1979). Chief executives define their own needs. *Harvard Business Review, 57*(2), 91 - 93.

Rogalewicz, M., Sika, R. (2016). Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering. *Management and Production Engineering Review, 7*(4), 97-108. doi:10.1515/mper-2016-0040

Saltz, J. S., Shamshurin, I. (2015). *Exploring the process of doing data science via an ethnographic study of a media advertising company*. Paper presented at the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA.

Saltz, J. S., Shamshurin, I. (2016). *Big data team process methodologies: A literature review and the identification of key factors for a project's success*. Paper presented at the 2016 IEEE International Conference on Big Data, Washington, DC, USA.

Saltz, J. S., Shamshurin, I., Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology, 68*(12), 2720-2728. doi:10.1002/asi.23873

Saunders, M., Lewis, P., Thronhill, A. (2016). *Research methods for business students* (7 ed.): Pearson Education Limited.

Schüritz, R., Brand, E., Satzger, G., Bischhoffshausen, J. (2017, June 5 - 10). *How to cultivate analytics capabilities within an organization? - Design and types of analytics competency centers.* Paper presented at the In Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal.

Sicular, S. (2012). No data scientist is an island in the ocean of big data. *Gartner*.

Verschuren, P., Doorewaard, H. (2015). *Het ontwerpen van een onderzoek* (5 ed.): Boom Lemma uitgevers.

Watson, H. J. (2014). Tutorial - Big data analytics concepts, technologies, and applications. *Communications of the Association for Information Systems, 34*, 1247 - 1268.

Yin, R. K. (2014). *Case study research and applications - design and methods*: Sage Publications Inc.

Bijlage 1 – Query's literatuuronderzoek

Query	Aantal	Relevant	Artikel	Jaar	Deelvraag
(TitleCombined:("data science projects"))	2	2	Predicting data science sociotechnical execution challenges by categorizing data science projects door Saltz, Jeffrey; Shamshurin, Ivan; Connors, Colin	2017	Type project (1)
			Successful Data Science Projects: Lessons Learned from Kaggle Competition door Al-Taie, Mohammed Zuhair; Salim, Naomie; Obasa, Adekunle Isiaka	2017	KSF (3)
(TitleCombined:("big data projects"))	17	3	Managing a Big Data project: The case of Ramco Cements Limited door Dutta, Debprotim; Bose, Indranil	2015	Methode project (2)
			Experience and reflection from China's Xiangya medical big data project door Li, Bei; Li, Jianbin; Jiang, Yuqiao	2019	Methode project (2)
			A Transdisciplinary Approach Supporting the Implementation of a Big Data Project in Livestock Production: An Example From the Swiss Pig Production Industry door Faverjon, C; Bernstein, A; Grutter, R	2019	Methode project (2) ; KSF (3)
(TitleCombined:("data science projects")) AND (TitleCombined:(type))	0				
(TitleCombined:("data science projects")) AND (TitleCombined:(sort))	0				
(TitleCombined:("big data projects")) AND (TitleCombined:(type))	0				
(TitleCombined:("big data projects")) AND (TitleCombined:(sort))	0				
(TitleCombined:("managing data science projects"))	0				
(TitleCombined:("manage data science projects"))	0				
(TitleCombined:("managing big data projects"))	0				
(TitleCombined:("manage big data projects"))	2	1	A decision-making approach based on fuzzy AHP-TOPSIS methodology for selecting the appropriate cloud solution to manage big data projects door Boutkhoum, Omar; Hanine, Mohamed; Agouti, Tarik	2017	Methode project (2)
(TitleCombined:("big data projects")) AND (manage)	6	3	Al eerder gevonden artikelen		
(TitleCombined:("big data projects")) AND (managing)	5	3	Al eerder gevonden artikelen		
(TitleCombined:("data science projects")) AND (manage)	1	1	Al eerder gevonden artikel		
(TitleCombined:("data science projects")) AND (managing)	1	1	Al eerder gevonden artikel		
(TitleCombined:("big data projects")) AND (TitleCombined:(methods))	1	0			
(TitleCombined:("data science projects")) AND (TitleCombined:(methods))	0				
(TitleCombined:("data science projects")) AND (methods)	1	1	Al eerder gevonden artikel		
("data science projects") AND (methods)	86	4	Modern data science for analytical chemical data – A comprehensive review door Szymańska, Ewa	2018	Methode project (2)
			Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering door Rogalewicz, Michał; Sika, Robert	2016	Methode project (2)
			Developing a business analytics methodology: A case study in the foodbank sector	2018	Methode project (2)

			door Hindle, Giles A; Vidgen, Richard		
			A review and future direction of agile, business intelligence, analytics and data science door Larson, Deanne; Chang, Victor	2016	Methode project (2)
(“big data projects”) AND (methods)	34	0			
(“data science projects”) AND (methodologies)	9	2	Al eerder gevonden artikelen		
(“big data projects”) AND (methodologies)	0				
(TitleCombined:(“data science projects”)) AND (TitleCombined:("critical success factors"))	0				
(TitleCombined:(“big data projects”)) AND (TitleCombined:("critical success factors"))	0				
(“data science projects”) AND ("critical success factors")	5	0			
(“big data projects”) AND ("critical success factors")	15	2	Critical success factors for Big Data adoption in the virtual retail: Magazine Luiza case study door Felix, BM; Tavares, E; Cavalcante, NWF	2018	KSF (3)
			Business intelligence for performance measurement: A case based analysis door Vallurupalli, Vamsi; Bose, Indranil	2018	KSF (3)
(“data science”) AND ("critical success factors")	7	0			
(“big data”) AND ("critical success factors")	38	0			

Tabel 4 – Query en zoekresultaten online Bibliotheek Open Universiteit

Query	Aantal	Relevant	Artikel	Jaar	Deelvraag
"data science project types"	2	1	Predicting data science sociotechnical execution challenges by categorizing data science projects door Saltz, Jeffrey; Shamshurin, Ivan; Connors, Colin	2017	Type project (1)
"data science project" type	506	5	Comparing Data Science Project Management Methodologies via a Controlled Experiment door Saltz, Jeffrey; Shamshurin, Ivan; Crowston, Kevin	2017	Methode project (2)
			Exploring the process of doing data science via an ethnographic study of a media advertising company door Saltz, Jeffrey; Shamshurin, Ivan	2015	Type project (1)
			Big data team process methodologies: A literature review and the identification of key factors for a project's success door Saltz, Jeffrey; Shamshurin, Ivan	2016	KSF (3)
			The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness door Saltz, Jeffrey	2015	Methode project (2)
			Successful Data Science Projects: Lessons Learned from Kaggle Competition door Al-Taie, Mohammed Zuhair; Salim, Naomie; Obasa, Adekunle Isiaka	2017	KSF (3)
"big data project" type	2040	7	Managing a Big Data project: The case of Ramco Cements Limited	2015	Methode project (2)

			door Dutta, Debprotim; Bose, Indranil		
			Big data project success - a meta analysis door Koronios, A; Gao, J; Selle, S.	2014	KSF (3)
			Towards A process view on critical success factors in big data analytics projects door Gao, J; Koronios, A; Selle, S	2015	KSF (3)
			4 al eerder gevonden artikelen		
"data science project" method	618		Al eerder gevonden artikelen		
"data science project" methodology	401		Al eerder gevonden artikelen		
"big data project" methodology	1410		Al eerder gevonden artikelen		
"big data project" "critical success factors"	117	1	Critical success factor categories for big data: A preliminary analysis of the current academic landscape door Eybers, S., Hattingh, M.J.	2017	KSF (3)
"data science project" "critical success factors"	30		Al eerder gevonden artikelen		

Tabel 5 – Query en zoekresultaten Google Scholar

Bijlage 2 – Overzicht literatuur

Nr	Titel	Auteur(s)	Jaar	Geciteerd[5]	Deelvraag
1	Predicting data science sociotechnical execution challenges by categorizing data science projects	Saltz, J.S., Shamshurin, I., Connors, C.	2017	16	Type project (1)
2	Exploring the process of doing data science via an ethnographic study of a media advertising company	Saltz, J.S., Shamshurin, I.	2015	19	Type project (1)
3	Comparing data science project management methodologies via a controlled experiment	Saltz, J.S., Shamshurin, I., Crowston, K.	2017	25	Methode project (2)
4	Synthesizing agile and knowledge discovery: case study results	Schmidt, C., Sun, W.N.	2018	5	Methode project (2)
5	A survey of data mining and knowledge discovery process models and methodologies	Mariscal, G., Marbán, Ó., Fernández, C.	2010	197	Methode project (2)
6	Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering	Rogalewicz, M., Sika, R.	2016	31	Methode project (2)
7	Managing a Big Data project The case of Ramco Cements Limited	Dutta, D., Bose, I.	2015	153	Methode project (2)
8	A review and future direction of agile, business intelligence, analytics and data science	Larson, D., Chang, V.	2016	189	Methode project (2)
9	Towards a process view on critical success factors in big data analytics projects	Gao, J., Koronios, A., Selle, S.	2015	45	KSF (3)
10	Big data team process methodologies: A literature review and the identification of key factors for a project's success	Saltz, J.S., Shamshurin, I.	2016	34	KSF (3)
11	Big data project success - a meta analysis	Koronios, A, Gao, J, Selle, S.	2014	19	KSF (3)
12	Critical success factor categories for big data: A preliminary analysis of the current academic landscape	Eybers, S., Hattingh, M.J.	2017	6	KSF (3)

Tabel 6 – Artikelen gebruikt voor theoretisch kader

[5] Volgens Google Scholar (november 2019)

Bijlage 3 – Literatuur appreciatie

Nr	Titel	Auteur(s)	Jaar
1	Predicting data science sociotechnical execution challenges by categorizing data science projects	Saltz, J.S., Shamshurin, I., Connors, C.	2017
	Het model om 4 typen data science projecten te creëren is gebaseerd op een studie bij 8 organisaties. Het is nodig om dit model bij meerdere organisaties te toetsen om na te gaan of de geïdentificeerde type projecten hier terugkomen. Daarnaast dient het model nog getoetst te worden op toepasbaarheid in de praktijk.		
2	Exploring the process of doing data science via an ethnographic study of a media advertising company	Saltz, J.S., Shamshurin, I.	2015
	Studie gebaseerd op onderzoek binnen één mediabedrijf en er voornamelijk op gericht om aanbevelingen te doen om data science projecten te verbeteren binnen dit bedrijf. De uitkomsten van dit artikel zouden moeten worden uitgebreid naar onderzoek in meer organisaties. De twee genoemde types data science projecten hebben geen uitgebreide getoetste basis.		
3	Comparing data science project management methodologies via a controlled experiment	Saltz, J.S., Shamshurin, I., Crowston, K.	2017
	Men introduceert een model om project methoden te vergelijken en bekijkt of de ene methode beter is dan de ander. Dit is in een gecontroleerde omgeving getoetst, maar met een aantal onderkende limitaties. Het onderzoek moet verder uitgebouwd worden voor representatieve resultaten in een data science omgeving.		
4	Synthesizing agile and knowledge discovery: case study results	Schmidt, C., Sun, W.N.	2018
	In dit onderzoek wordt er een raamwerk gebouwd door het samenvoegen van KDD, CRISP-DM en agile kenmerken op basis van literatuur. Het raamwerk is getoetst in een case study binnen één organisatie. De auteurs geven aan dat het voorgestelde raamwerk getoetst moet worden in meerdere organisaties.		
5	A survey of data mining and knowledge discovery process models and methodologies	Mariscal, G., Marbán, Ó., Fernández, C.	2010
	Een beschrijvend onderzoek op basis van literatuur naar procesmodellen en methoden om big data projecten te doorlopen, waarbij de methoden en processen naast elkaar worden gezet en worden vergeleken op de te doorlopen fases. Op basis hiervan is een verfijnd data mining proces beschreven, waarbij de nadruk is gelegd op de subprocessen. Er is nog geen concrete methode met life cycle beschreven. Tevens is het verfijnde proces niet in de praktijk toegepast.		
6	Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering	Rogalewicz, M., Sika, R.	2016
	Onderzoek gericht op één werkveld, waarbij men een aantal data mining methoden bespreekt op basis van (wetenschappelijke) literatuur. Er wordt geen methode van onderzoek beschreven. Het artikel geeft inzicht in de meest gebruikte taken binnen data mining methoden. Het artikel is voornamelijk beschrijvend of informatief van aard.		
7	Managing a Big Data project The case of Ramco Cements Limited	Dutta, D., Bose, I.	2015

	Het artikel richt zich op een raamwerk gebaseerd op wetenschappelijke literatuur. Het raamwerk beschrijft een roadmap voor het succesvol uitvoeren van een big data project. Zoals de titel aangeeft is het raamwerk getoetst binnen één organisatie. De methode van toetsing is uitgebreid beschreven, maar het raamwerk dient in meerdere organisaties in andere werkvelden getoetst te worden.		
8	A review and future direction of agile, business intelligence, analytics and data science	Larson, D., Chang, V.	2016
	Het artikel richt zich met name op het samenstellen van een raamwerk bestaande uit een business intelligence life cycle met daarbij agile principes toegepast. Er is geen methode van onderzoek beschreven, maar het raamwerk is gebaseerd op wetenschappelijke literatuur. Het voorgestelde raamwerk is niet getoetst.		
9	Towards a process view on critical success factors in big data analytics projects	Gao, J., Koronios, A., Selle, S.	2015
	Onderzoek dat gebaseerd is op niet alleen wetenschappelijke literatuur, maar ook blogs om alle KSF omtrent big data analytics in beeld te krijgen. KSF'n zijn ingedeeld in fases van het "business analysis process" van Sicular (2012). De uitgebreide lijst KSF'n is niet getoetst in de praktijk, zodat niet bekend is wat de waarde voor succes bij een big data project.		
10	Big data team process methodologies: A literature review and the identification of key factors for a project's success	Saltz, J.S., Shamshurin, I.	2016
	Uitgebreide literatuurstudie om KSF'n te identificeren, maar de lijst met KSF'n dient nog verfijnd, eventueel geprioriteerd en gevalideerd te worden in meerdere case studies.		
11	Big data project success - a meta analysis	Koronios, A, Gao, J, Selle, S.	2014
	Kwalitatief onderzoek door KSF'n te identificeren op basis van case studies, waarbij een aantal KSF'n zijn gebaseerd op (niet betrouwbare) secundaire data. KSF'n zijn ingedeeld in fases van het "business analysis process" van Sicular (2012). De uitgebreide lijst KSF'n is niet getoetst in de praktijk, zodat niet bekend is wat de waarde voor succes bij een big data project.		
12	Critical success factor categories for big data: A preliminary analysis of the current academic landscape	Eybers, S., Hattingh, M.J.	2017
	Het artikel identificeert KSF'n op basis van wetenschappelijke literatuur en met een kwantitatieve methode worden de KSF'n ingedeeld in categorieën. De methode van onderzoek is duidelijk beschreven. Men geeft aan dat de onderzoeksresultaten verder onderzoek nodig hebben, bv. het vergelijken van de KSF'n met ander onderzoek en of de KSF'n toepasbaar zijn. Het voorgestelde raamwerk is niet getoetst.		

Tabel 7 – Beoordeling kwaliteit artikel

Bijlage 4 – Kritieke succesfactoren

Nr	KSF en *fase*
	Business phase
1	Identifiable business value
2	Clear and manageable project scope
	Data phase
3	Identification and access to needed data sources
4	Combine different data sets
5	High data quality
6	Data security and privacy
	Analysis phase
7	Innovative analysis tools
8	Adequate hardware
9	Analytical skillset
10	Technical skillset
11	Integration of new solutions
12	Fast delivering of new results
13	Cloud-based solutions
14	Flexible IT-structure
15	Visualization
16	Virtualization
17	Adapt architectural principles
	Implementation phase
18	Information strategy for big data
19	Big data as strategic instument
20	Interpretation of analytical results
	Measurement phase
21	Clear project goal with deadline
22	Measureable outcome
	Overall phase
23	Top management support
24	Multidisciplinary teams
25	Independent business unit
26	Iterative process model
27	Outsourcing

Tabel 8 – KSF Gao et al (2015)

Nr	KSF en *categorie*
	Data
1	Data & data quality management / ownership
2	Data integration & security
3	Unstructured/structured data
4	Representativeness of data

5	Document collection/access to sources
	Governance
6	Management priority / sponsorship / support
7	Big data strategy alignment (with organization's vision)
8	Project management process defined
9	Well defined organizational structure
10	Performance management
11	Data protection and privacy by design
12	Culture of being data-driven
	Process
13	Close collaboration between IT and business
14	Communication about the data and initiatives
15	Flexibility and agility, with freedom for experimentation
16	Focus on change management
17	Project difficulty explored and communicated
18	Clarity of project deliverables (clear or ambiguous)
	Objectives
19	Focus on small projects and known questions
20	Specified business case
21	Feasibility study
22	Skill gap analysis
23	Well defined scope - that understood by the team
24	Measurable project outcome
	Team
25	Development of skills / training
26	People skills & ability to self-organize when needed
27	Data science, technology, business & management skills
28	Multidisciplinary team (i.e., across different departments)
29	Stakeholder coordination / shared understanding
	Tools
30	Investment in IT infrastructure, technology & tools
31	Investment in data sources & data storage
32	Reporting and visualization technology
33	Discovery technology

Tabel 9 – KSF Saltz & Shamshurin (2016)

Bijlage 5 – Interviewprotocol

Interview wordt bij voorkeur gehouden op de werklocatie van de te interviewen personen.

1. Afspraak inplannen (bij voorkeur telefonisch) en eventueel context van het onderzoek (digitaal) toesturen.

   Interview uitvoeren
2. Inleiden gesprek
   Voor mijn afstudeeronderzoek aan de Open Universiteit (Master Business Process Management & IT) heb ik de opdracht te onderzoeken hoe data science projecten worden uitgevoerd. Op dit moment is data science een "booming" onderwerp.
   In mijn onderzoek wil ik meer inzicht verkrijgen in de manier waarop data science projecten succesvol kunnen worden uitgevoerd. In de wetenschappelijke literatuur wordt gesproken over kritieke succesfactoren (KSF'n), waar ik later op in ga. Maar ook over type data science projecten en methoden om data science projecten te managen. Ik ben benieuwd of ik een relatie kan ontdekken tussen de KSF'n, de type projecten en de methoden.
   Door mijn gesprek met u hoop ik de informatie te verzamelen die ik nodig heb om mijn onderzoek af te ronden.
   Heeft u hier vragen over?
3. Het gesprek zal vertrouwelijk behandeld worden en de gegevens (van geïnterviewde en organisatie) zullen geanonimiseerd worden. Daarnaast heeft u het recht om op elk moment vragen te weigeren. Tijdens het interview zal ik notities maken en van het gesprek wordt een audio opname gemaakt, zodat ik makkelijker een verslag van het gesprek kan maken.
4. Om de afspraken omtrent het interview vast te leggen, heb ik u vooraf al het consent formulier toegestuurd. Als u het interview wilt voortzetten, dan kunnen we het consent formulier ondertekenen.

Dan zal ik nu de opname starten en beginnen met het interview.

Q1a. Kunt u uw rol of functie beschrijven in het data science team?
Q1b. Wat verstaat u onder data science binnen uw organisatie?
Q1c. Wat verstaat u onder een data science project?
Q1d. Kunt u uw ervaring beschrijven in data science projecten?

Onderzoeksvraag 1: Wat voor typen data science projecten kunnen worden onderscheiden?
Q2a. Worden er data science projecten binnen uw afdeling of door uw team uitgevoerd?
Q2b. En kunt u voorbeelden geven van data science projecten?
Q2c. Kunt u deze data science projecten nader beschrijven?
Q2d. In welke typen of categorieën zijn deze data science projecten volgens u in te delen?

Als er geen typen data science projecten herkend worden door de geïnterviewde:
Q2e. Herkent u de vier typen data science projecten die beschreven zijn in de literatuur (zie matrix en zo nodig beschrijven)?
Q2f. Zijn de data science projecten van de organisatie toe te passen in de matrix uit de literatuur?

Nu ik meer inzicht heb in welke data science projecten er uitgevoerd worden, wil ik de stap maken naar hoe data science projecten worden uitgevoerd. De komende vragen gaan hierop in.

Onderzoeksvraag 2: Welke methoden zijn er beschikbaar voor het managen van data science projecten?
Q3a. Wie initieert data science projecten? En wie is er vervolgens verantwoordelijk voor?
Q3b. Hoe worden data science projecten binnen uw team gemanaged?
Q3c. Welke methode voor het managen van een project wordt toegepast?
Q3d. Waarom wordt deze methode toegepast?
Q3e. Hoe wordt deze methode toegepast?
Q3f. Hoe ziet de samenstelling van het data science team in een project er uit?
Q3g. Als u een bepaald type project uitvoert (neem als voorbeelden de projecten genoemd in Q2b), welke methode wordt dan toegepast om het project te managen? En waarom?

Na de methoden die gebruikt worden bij het uitvoeren van data science projecten besproken te hebben, wil ik nu in gaan op de mate van succes in projecten.

Onderzoeksvraag 3: Welke kritieke succesfactoren zijn van invloed op data science projecten?
Q4a. Kunt u een voorbeeld geven van een succesvol project?
Q4b. Wat ging er goed in het project?
Q4c. Wat was een kritieke factor in het project om het project te laten slagen? Of waren er meer factoren volgens u?
Q4d. Op welk gebied kwamen deze factoren voor: technisch, organisatorisch, etc?
Q4e. Wat ging er mis in het project?
Q4f. Wat had volgens u beter gekund in het project? Op welk gebied: data, governance, process, objectives, team, tools?
Q4g. Zijn er ook projecten mislukt? Zo ja, waarom is dat project mislukt (en is het andere project geslaagd)?
Q4h. Wordt er in de data science projecten gestuurd op basis van kritieke succesfactoren?
Zo ja, welke KSF'n (doorvragen met Q4i en Q4j)? (Zo nee, dan door naar vraag Q4l)
Q4i. Verschillen deze KSF'n per project en wat is de reden hiervoor? (in welke mate verschillen deze KSF per project?)
Q4j. Verschillen deze KSF'n per fase in het project en wat is de reden hiervoor?
Q4k. Als u een bepaalde methode toepast bij het uitvoeren van een project (neem als voorbeelden de methoden genoemd in Q3b), welke KSF'n worden dan toegepast om het project te managen en te laten slagen? En waarom?

Q4l. Kunt u aan de hand van deze lijst met KSF'n aangeven welke 10 KSF'n volgens u het meest belangrijk zijn om een data science project succesvol te managen? U kunt in het opmerkingenveld eventueel uw antwoord nader toelichten.

Hierbij zijn we aan het einde gekomen van het interview.

Q5. Heeft u tenslotte nog vragen/opmerkingen/aanvullingen die niet ter sprake zijn gekomen tijdens het gesprek?

Dan wil ik u graag bedanken voor uw tijd en de informatie die u gegeven heeft. Dat helpt mij enorm bij het uitvoeren van mijn onderzoek. Ik zal een verslag maken van dit gesprek en dat aan u toesturen. Het verzoek aan u om dit verslag te controleren om na te gaan of uw antwoorden door

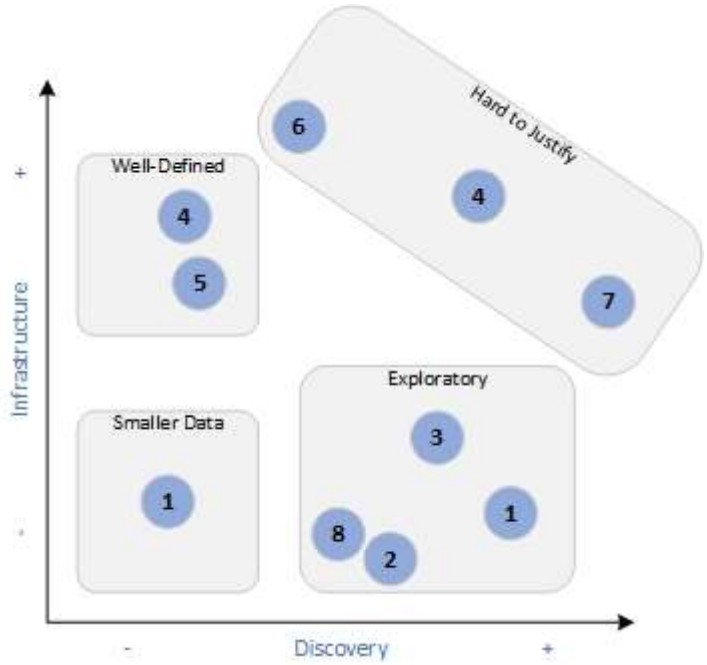
mij correct zijn weergegeven. Kunt u mij dit vóór (*datum overeenkomen*) laten weten of het verslag met opmerkingen aan mij terugsturen?
Bent u ook geïnteresseerd in de uitkomsten van het onderzoek? Zo ja, dan zal ik u deze toesturen.
Mocht u nog vragen hebben, dan kunt u mij bereiken op “mailadres” en “telefoonnummer”.

Na interview

5. Verslag/transcript schrijven en naar de geïnterviewde sturen voor feedback.

Het door de geïnterviewde becommentarieerde verslag verwerken, zodat het gereed is voor analyse.

Matrix type projecten te gebruiken bij vragen Q2e en Q2f:



Figuur – Vier typen data science projecten (Saltz et al, 2017)

- Moeilijk te rechtvaardigen: project heeft geen helder doel, maar vereist vooraf een grote investering. Moeilijk om ondersteuning vanuit de organisatie te krijgen.
- Verkennend: project heeft geen helder doel, dus makkelijker om zaken te proberen. Lage kosten door minder vereiste infrastructuur. Een project zonder helder doel.
- Duidelijk gedefinieerd: project heeft een duidelijk doel, maar vergt een grote investering. Vooraf is namelijk te rechtvaardigen dat de investering nut heeft.
- Weinig data: project met een duidelijk doel, maar vergt een kleine investering in infrastructuur.

Door de respondent aan te geven welke de 10 meest belangrijke KSF’n zijn met een eventuele toelichting (vraag Q4l):

Nr	**KSF en *categorie***	**Belangrijk**
	Data	
1	Data & data quality management / ownership	

2	Data integration & security	
3	Unstructured/structured data	
4	Representativeness of data	
5	Document collection / access to sources	
	Governance	
6	Management priority / sponsorship / support	
7	Big data strategy alignment (with organization's vision)	
8	Project management process defined	
9	Well defined organizational structure	
10	Performance management	
11	Data protection and privacy by design	
12	Culture of being data-driven	
	Process	
13	Close collaboration between IT and business	
14	Communication about the data and initiatives	
15	Flexibility and agility, with freedom for experimentation	
16	Focus on change management	
17	Project difficulty explored and communicated	
18	Clarity of project deliverables (clear or ambiguous)	
	Objectives	
19	Focus on small projects and known questions	
20	Specified business case	
21	Feasibility study	
22	Skill gap analysis	
23	Well defined scope - that understood by the team	
24	Measurable project outcome	
	Team	
25	Development of skills / training	
26	People skills & ability to self-organize when needed	
27	Data science, technology, business & management skills	
28	Multidisciplinary team (i.e., across different departments)	
29	Stakeholder coordination / shared understanding	
	Tools	
30	Investment in IT infrastructure, technology & tools	
31	Investment in data sources & data storage	
32	Reporting and visualization technology	
33	Discovery technology	

Tabel – KSF Saltz & Shamshurin (2016)

Toelichting bij keuze kritieke succesfactoren

Bijlage 6 – Ingevulde lijst KSF door respondenten

Nr	KSF en *categorie*	Int1	Int2	Int3	Int4	Aantal
	Data					
1	Data & data quality management / ownership	x	x		x	4
2	Data integration & security					
3	Unstructured/structured data					
4	Representativeness of data	x		x	x	3
5	Document collection / access to sources					
	Governance					
6	Management priority / sponsorship / support	x		x	x	3
7	Big data strategy alignment (with organization's vision)					
8	Project management process defined					
9	Well defined organizational structure					
10	Performance management					
11	Data protection and privacy by design					
12	Culture of being data-driven	x	x	x		3
	Process					
13	Close collaboration between IT and business	x	x	x	x	4
14	Communication about the data and initiatives					
15	Flexibility and agility, with freedom for experimentation		x	x		2
16	Focus on change management				x	1
17	Project difficulty explored and communicated					
18	Clarity of project deliverables (clear or ambiguous)	x		x	x	3
	Objectives					
19	Focus on small projects and known questions		x			1
20	Specified business case	x			x	2
21	Feasibility study		x			1
22	Skill gap analysis					
23	Well defined scope - that understood by the team			x		1
24	Measurable project outcome			x	x	2
	Team					
25	Development of skills / training			x		1
26	People skills & ability to self-organize when needed			x		1
27	Data science, technology, business & management skills	x			x	2
28	Multidisciplinary team (i.e., across different departments)	x				1
29	Stakeholder coordination / shared understanding		x			1
	Tools					
30	Investment in IT infrastructure, technology & tools				x	1
31	Investment in data sources & data storage		x			1
32	Reporting and visualization technology	x				1
33	Discovery technology					

Tabel 10 – KSF Saltz & Shamshurin (2016) ingevuld door respondenten met toelichting

Toelichting bij keuze kritieke succesfactoren

Int1:
4 – anders heeft een gebouwd model geen nut.
6 – van belang binnen de organisatie voor de steun en vraag ernaar (bestaansrecht)
Int2:
Voor wat betreft de data: deze is er wel maar de kwaliteit en kwantiteit laten te wensen over. De combinatie van voldoende goede data en de domein kennis (betekenis van de data) zijn voor mij een belangrijke eerste stap.
Verder moet de hele organisatie (dus ook de echte ijzervreters) zich bewust worden van het belang en moeten er handvatten worden aangereikt zodat ze meer datagedreven kunnen gaan acteren.
Er moet een goede balans zijn tussen bekende uitdagingen en uitdagingen die voor de organisatie nog buiten het gezichtsveld liggen maar noodzakelijk zijn voor verdere ontwikkeling. Elke actie of werkpakket moet de menselijke maat hebben (dus klein qua omvang).
Datascience is nu een hype. Veel mensen hebben er een torenhoge verwachting bij. Het managen van deze verwachting bij de stakeholders en opdrachtgevers en natuurlijk ook de gebruikers is van belang om niet de dynamiek te verliezen.
Databronnen toegankelijk maken en beschikbaar krijgen zijn zoals al gemeld nu belangrijke succesfactoren hierbij is de noodzakelijke infrastructuur onontbeerlijk.
Wat ik een beetje mis in de lijst is de zachte kant van het gebruik van datascience toepassingen. Denk hierbij aan de wil om het te gebruiken onder extreme omstandigheden, het vertrouwen in de geboden ondersteuning etc…
Int3:
Representativeness of data: als de data waarmee je ontwikkelt/experimenteert niet representatief is, zal je model in de praktijk niet dat doen wat je wilt. Wat je erin stopt krijg je er ook uit. Stop je er rommel in, dan geeft een model ook rommel terug. In sommige sectoren is momenteel nog geen goede data voor handen om mee te experimenteren, dan moet je naar oplossingen zoals synthetische data gaan kijken, maar in mijn ervaring is dat nog nooit goed uitgepakt.
Management priority / sponsorship / support: wat als je project slaagt? Dan wil je er natuurlijk een klap op geven, en moet er vaak extra geld komen om het door te ontwikkelen, te beheren, etc. Als er geen support is, is er ook geen geld.
Culture of being data-driven: in alle lagen van de organisatie moet deze cultuur gaan ontstaan, ook al is het voor sommigen misschien een ver-van-je-bed-show. Bijvoorbeeld bij inkopers moet op het netvlies staan dat bij inkoop van nieuwe voertuigen ook toegang tot de sensordata meegekocht moet worden (dit is nog lang niet altijd het geval), en bij degenen die het voertuig besturen moet bekend zijn dat die data belangrijk is, en dat sensoren dus niet zomaar uitgezet, verwijderd, etc. moeten worden. Alleen als iedereen meedoet is de kwaliteit van de data voldoende zodat er mooie dingen mee gedaan kunnen worden.
Close collaboration between IT and business: hier hebben we het al uitgebreid over gehad
Flexibility and agility, with freedom for experimentation: je gaat hoe dan ook tegen problemen aanlopen en flexibiliteit stelt je in staat voor de diversiteit aan problemen een oplossing te vinden.
Clarity of project deliverables (clear or ambiguous): Wanneer is een project succesvol? Deze vraag kan alleen beantwoord worden als de deliverables duidelijk zijn.
Well defined scope - that understood by the team: Het is soms heel leuk een uitstapje te maken, en 1 klein onderdeeltje heel goed uit te zoeken. Nerds vinden misschien wel niets leuker. Maar, soms leidt dit teveel af van het grotere doel, namelijk project X op manier Y in 4 maanden voltooien. De scoping helpt hierbij.

Measurable project outcome: zelfde als clarity of project deliverables. Wat is je deliverable en hoe beoordeel je of je project succesvol is? Ontzettend belangrijk.
Development of skills / training: Dit vakgebied is elke dag in beweging, en er gebeurt meer dan je bij kunt leren. Als data scientist moet je vooral snel nieuwe dingen kunnen bijleren. Als je dit niet doet ben je straks niet meer relevant.
People skills & ability to self-organize when needed: bij het Datalab zijn we naast data scientist ook scrum master, want die taken doen we zelf. We zijn zelf verantwoordelijk voor het contact met de business, en daar heb je skills voor nodig.
Int4:
1 - Toegang tot de data benodigde data is essentieel. Daarbij belangrijk om dus goede contacten te hebben met eigenaar van de data. Ook handig om evt. terug te kunnen koppelen over de kwaliteit van de data zodat de eigenaar kan kijken of hij dat kan verbeteren.
4 - Als de data niet representatief is zijn praktische toepassingen van een model beperkt of onmogelijk.
6 - Sponsorschap vanuit de organisatie is erg belangrijk. Tijdens een project kom je altijd (menselijke) obstakels tegen binnen de organisatie en die wil je kunnen doorbreken met behulp van iemand die hoger in de boom zit.
13 - Continu contact met de business (eindgebruiker) vergroot de kans dat het project echt waarde gaat toevoegen voor de organisatie. Ook helpt het met draagvlak.
16 - Gerelateerd aan punt 13. De eindgebruiker moet meegenomen worden in de ontwikkeling en voorbereid worden op een nieuwe manier van werken obv het eindproduct of aanbevelingen.
18 - Duidelijke deliverables zijn nodig om focus te kunnen houden
20 - Gerelateerd aan punt 6. Een goede business case vergroot de kans dat het project succesvol wordt en ook tot een goed einde kan worden gebracht (door draagvlak binnen de organisatie).
24 - Zonder meetbaar resultaat zou men binnen de organisatie van project naar project gaan, zonder dat ooit duidelijk wordt of een afgerond project daadwerkelijk oplevert wat het beloofde. Belangrijk om daar kritisch in te zijn. Zonder meetbaar resultaat is het ook onmogelijk om gericht verbeteringen door te voeren.
27 - Al deze skills zijn belangrijk om een project succesvol te maken. Technische mensen voor de techniek, business mensen voor procesmatige zaken en zorgen dat de technische mensen hun efforts op de juiste zaken richten.
30 - Vaak een randvoorwaarde, de IT omgeving moet geschikt zijn om het project uit te voeren (bijv zwaar model trainen, maar ook überhaupt toegang tot data) en een eindproduct op te leveren.